University of Alberta

Hardwired Freedom: Illusion as a Vehicle for Moral Responsibility

by

Bartlomiej Lenart

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Arts

Department of Philosophy

Edmonton, Alberta
Fall 2007

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence
ISBN: 978-0-494-33137-8
Our file Notre référence
ISBN: 978-0-494-33137-8

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

This essay explores how Peter Strawson's notion of reactive attitudes enriches Drew McDermott's illusionist view regarding free will by supplementing McDermott's theory with a compatibilist framework that preserves moral responsibility within McDermott's deterministic and mechanistic vision of the mind. In addition to providing McDermott's theory of mind with a robust moral framework, the reactive attitudes are fortified by being grounded in McDermott's notion of a self-model. The reactive attitudes and McDermott's conception of self-modelling (along with the illusionism inherent in McDermott's self-model view) reinforce each other. This essay also explores the possibility of the compatibility of freedom and quantum indeterminacy as found in Robert Nozick's contemplations in his *Philosophical Explanations*. Although such a view is capable of accounting for genuinely open alternatives, such indeterminacy can, at most, amount to some (often quite limited) degree of self-formation and/or re-formation, which substantially narrows the scope of human freedom.

Acknowledgments

First and foremost, I owe a great debt of gratitude to my thesis supervisor, Professor Wesley Cooper, for his steadfast support, enthusiasm, encouragement, patience, suggestions, and insight throughout the entire process of conceptualizing and writing this thesis.

I also offer many thanks for the insightful comments, questions, and suggestions of my committee, Professors Renée Elio, Bruce Hunter, and Adam Morton, who have inspired further contemplation of the subject matter.

I thank my family, Andrzej, Urszula, and Kamil Lenart, for continuously supporting me in my decision to study philosophy and for always being a source of encouragement, motivation, and comfort.

I am fortunate to have many fine friends willing to take the time to listen to my thoughts and discuss issues of free will with me. Special thanks go to Anahi Johnson for her discerning intellect, ardent discussions, and the countless hours of editing, questioning, and analyzing, Ela Przybylo for her probing intellectual disagreement with me and for her comforting and reassuring encouragement, and Lisa Rusnak for her unstinting willingness to listen to my ideas and converse with me about them.

Finally, I would like to extend my thanks to several open-minded friends and fellow graduate students, Yual Chiek, Charles Rodger, and Jason Taylor, for their thought-provoking and intellectually stimulating views and conversations.

Table of Contents

Introduction 1

Chapter One
Computationalism and Freedom of the Will

1.1	Mechanism and Computationalism	5
1.2	Computers and Free Will	7
1.3	What is a Self-Model?	13
1.3.1	Virtual Selves	16
1.3.2	The Phenomenal Experience of Self ("Mine," "Me," and "Myself")	19
1.3.3	Transparency	20

Chapter Two
McDermott on Self-Modelling, Qualia, and Free Will

2.1	Self-Modelling and Deliberation	26
2.2	Qualia	28
2.3	Symbols and Semantics	33
2.4	Qualia and Free Will	39
2.5	Non-Moral Judgment and Natural Human Reactions	47

Chapter Three
A Naturalistic Approach to Responsibility

3.1	Reactive Attitudes	54
3.2	Illusion	61
3.3	Reactive Attitudes and Illusionism	66
3.4	Frankfurt's Compatibilism	68
3.5	Reactive Attitudes (Objections and Replies)	76
3.6	Hardwired Freedom and Responsibility	88

Chapter Four
Looking for Freedom in Other Places

4.1	On Second Thought…	93
4.2	Choice and Indeterminism According to Nozick	95
4.3	The Problem of Arbitrariness	99
4.4	Finite Beings, Finite Freedom?	108
4.5	Some Benefits of Quantum Randomness: Alms for the Freeless	114

Conclusion 118

Bibliography 119

Appendices 123

Introduction

When addressing his teacher St. Augustine of Hippo, Euodius says: "I am deeply troubled by a certain question: how can it be that God has foreknowledge of all future events, and yet that we do not sin by necessity" (St. Augustine 391, 260)? Euodius' question, even after sixteen centuries, continues to be a troubling one. Today, we may wish to phrase the concern slightly differently.

The philosophical problem of free will, as addressed in this essay, can be stated as follows: modern biology (and science in general) describes human beings as machines or mechanisms that are influenced by the laws of nature to the same extent as simple organisms or fully inanimate objects. All the while, however, human beings have the experience of free will, which grants them special status in what they perceive to be a mechanical universe. This notion of free will has become a foundation of sorts for our judgments about human actions; the notion of free will is central to our evaluation of human actions as praiseworthy or blameworthy, right or wrong, etc. Assuming that science is correct in its assertions about the nature of the world, human bodies, and the mind, how can we reconcile the notion of free will, which is quite vital to our human affairs in the social realm as well as to our more philosophically inclined interests, with the mechanistic, physicalist story, which science has unfolded? What is so special about some events that they can be evaluated, judged, praised, and blamed? I am deeply troubled by the problem of free will.

It seems to me that what differentiates agents from rocks is the mind. If there is to be an answer to the problem of free will, then such an answer should make

mention of the mind or at least say something about some of the things it does. If the mind is the key to our freedom, then we may continue to remain deeply troubled for some time to come since the mind is likely one of the greatest unsolved mysteries facing science today.

Having said this, however, my curiosity compels me to stumble across the many problems and questions that riddle the issues concerning freedom of the will. Perhaps the best strategy to adopt when one is unfamiliar with one's surroundings is to take small exploratory steps and not venture too far out into the vast unknown. Thus, I shall be limiting myself to one particular view of the mind (and one particular proponent of this view) in my explorations.

In contemporary mind science, the computer has become a powerful analogy for the structure and functions of the human mind. Researchers in the field of Artificial Intelligence (AI) have taken this analogy quite seriously and many, in fact, posit that the human brain is nothing more than a biological computer. Drew V. McDermott, an AI researcher, in his book *Mind and Mechanism*, develops a computationalist approach to the mind. He sketches a computational notion of a self-model, which attempts to solve various problems including the problem of phenomenal consciousness. McDermott's self-model also has important implications for the question of free will.

This essay concerns itself with Drew McDermott's view regarding free will, which arises out of his computational model of mind, and more specifically, out of the notion of the self-model. One problem McDermott's view faces is the issue of accountability or moral responsibility. McDermott's vision of the mind is a

deterministic one and thus, if having the ability to do otherwise is necessary for notions of accountability and responsibility (as many philosophers have argued), then genuine freedom of the will is required for the practices of praising and blaming (in the moral sense) to be meaningful. This essay is, in effect, an enrichment project; it is an attempt to supplement McDermott's view with a compatibilist theory that allows for the notion that freedom of the will is an illusion and yet lays the foundations for moral responsibility.

My main purpose will be to (i) explore whether the computationalist view can account for our experience of freedom (which I think it can even though I do not think McDermott's view accounts for metaphysical freedom, but merely for the illusory belief in free will) and (ii) to attempt to buttress McDermott's vision of free will with a suitable account of moral responsibility. To this end, I shall explore Peter Strawson's notion of reactive attitudes. And finally, I wish to investigate the possibility of a libertarian freedom in a purely physical universe. At this point I will follow in Robert Nozick's footsteps and will consider the possibility of the compatibility of quantum indeterminacy with metaphysical freedom.

Chapter one is primarily an expository chapter intended to introduce and clarify the notion of a self-model. Here, Thomas Metzinger's view of self-modelling will serve as a suitable background to the notion of self-modelling McDermott utilizes.[1] The discussions of chapter two include the relationship between

[1] My focus, however, will solely be on McDermott's version of self-modelling and thus, I will not concern myself with Metzinger's other philosophical commitments. In other words, Metzinger's analysis of self and consciousness is important only insofar as it contributes to my understanding of McDermott's notion of self-modelling. Any differences between Metzinger and McDermott are ignored. Such omission is justified, I think, by the similarities (between the two thinkers) I wish to draw on, by my sole focus on McDermott (as this essay is an enrichment of McDermott's view), and by the fact that a certain amount of simplicity will contribute to rather than detract from my thesis,

deliberative processes and self-modelling, McDermott's understanding of qualia, and the connection between qualia and free will. Chapter two closes with an examination of J. J. C. Smart's notion of non-moral judgments. The promised enrichment of McDermott's view of freedom by means of the Strawsonian notion of reactive attitudes is found in chapter three where illusionism about free will is also discussed in greater detail (more specifically, the views of the psychologist Daniel M. Wegner and the philosopher Saul Smilansky will prove to be quite useful). Also, chapter three explores the possibility of the separation between metaphysical freedom and moral responsibility (to this end, I employ Harry Frankfurt's argument against the principle of alternative possibilities). The final chapter explores the possibility of metaphysical freedom in light of quantum indeterminacy.

---

which is concerned with the problem of free will and moral responsibility rather than the problem of consciousness.

Chapter One
Computationalism and Freedom of the Will

1.1
Mechanism and Computationalism

"The idea that the mind is a natural mechanism derives from thinking of nature itself as a kind of mechanism"[2] (Crane 2003, 2). The thesis of determinism, which is implied by universal mechanism, the view that the universe is best understood as a completely mechanical system, may not be as prominent today as it once was. The truth of universal determinism has been put into doubt especially in light of developments in the field of quantum mechanics. However, anthropic mechanism, the thesis that everything about human beings can, in principle, be explained in mechanistic terms seems, at least in some prominent circles, to be alive and well.[3]

Vitalism, the idea that what distinguishes living entities from lifeless matter is a mysterious vital force (an *élan vital*), has been discarded as a useless notion incapable of explaining the origin of life and the beginnings of mind. The search for the source of consciousness has led many exceptional thinkers to compare the mind to the newest and most complicated machinery of their day. John Searle's comment serves as an example as well as a warning not to be too hasty in drawing analogies or accepting the latest metaphors as true explanations of the mind:

> Because we do not understand the brain very well we are constantly tempted to use the latest technology as a model for trying to understand it. In my childhood we were always assured that the brain was a telephone switchboard. ('What else could it be?') I was amused to see that Sherrington, the great British neuroscientist, thought that the brain worked like a telegraph system. Freud often compared the brain to hydraulic and electro-magnetic systems. Leibniz compared it to a mill, and I am told that some of the ancient Greeks thought the brain

[2] The Oxford Dictionary of Science defines mechanics as "[t]he study of the interactions between matter and the forces acting on it" (Isaacs, Daintith, Martin 1999, 490).
[3] Drew McDermott's computationalist view is an example.

> functions like a catapult. At present, obviously, the metaphor is the digital computer. (Searle 1984, 44)

Although computationalism, which insists that cognition is a form of information processing, is much more sophisticated in its explanations than the views that described the mental in terms of telephone switchboards, hydraulic systems, mills, and catapults, it must be remembered that the digital computer (as well as the more recent connectionist systems) are at best approximations or metaphors.[4] Having said this, however, one should not dismiss computationalism without a second thought. After all, many great discoveries of the past defied (and continue to defy) commonsense. The earth, for example, may be in motion, but as far as we are concerned the sun "sets" and "rises" and once in a while we catch ourselves realizing that the rising and setting of the sun is just an illusion. Of course, it would be a mistake, on the other hand, to reject all commonsense notions just because some proved to be inaccurate.[5]

Computationalism, with its many opponents, must stand the test of various criticisms. In this essay, I shall concern myself with only one worry, namely the

---

[4] McDermott recognizes the fact that his view is just a sketch. He states that the difficulty with defining consciousness has led some philosophers to propose that there is no such thing as consciousness while others, notably Colin McGinn, do not think we have the tools we need to explain the mind (McGinn is a pessimist in that he does not believe we ever will have the tools). McDermott, on the other hand, thinks of himself as an optimist in that he both thinks there is such a thing as consciousness and that we do or at least will have the necessary tools to explain the mind. However, he continues: "The idea is to replace the concept of consciousness with more refined (more scientific?) concepts, much as happened with concepts like 'energy' and 'mass' in past scientific revolutions. It seems plain that a full understanding of the mind will involve shifts of this kind. If we ever do achieve fuller understanding (which the pessimists doubt), any book written before the resulting shift, including this one [McDermott's own book *Mind and Mechanism*] will no doubt seem laughably quaint" (McDermott 2001, 20). In the same breath, McDermott adds: "However, we can't simply wait around for this to happen. We have to work on the problems we see now, using the tools at hand" (McDermott 2001, 20).

[5] I shall argue that, on McDermott's view, the commonsense notion of free will is, in fact, an illusion, which has no place in our ontology whereas our commonsense notion of moral responsibility continues to hold even in the absence of metaphysical freedom.

problem of free will and the implications on our commonsense notion of moral accountability.

1.2
Computers and Free Will

According to Eric Dietrich, "[t]he term 'computation' denotes a step-by-step, mechanical process with a definite beginning and a definite end where each step (called a state) and each transition from one state to the next is finitely and unambiguously describable and identifiable" (Dietrich 1994, 7). A mechanical system like a computer must, at its most basic structural level, be understood as a causal system. Computers are commonly thought of as syntactic engines, but as David Cole insightfully explains, it is still misleading to think of computers as syntax manipulators. "CPUs do not follow syntactic rules. They cannot literally manipulate 1s and 0s" (Cole 1994, 143). Computers, according to Cole, do not understand syntax any more than they understand semantics. Computers "are causal systems that have causal properties that mirror—are isomorphic with—syntactic transformations" (Cole 1994, 144).

It seems appropriate to view causal systems, and more specifically computational systems, as deterministic entities. Given a certain input i_1, a machine in an initial state s_1 (if working properly and if programmed to perform a certain function f) will produce some output o_1. The processes responsible for the manipulation of symbols, which eventually result in o_1 given i_1, are causal processes, which adhere to the laws of nature. Moreover, given the same initial settings (being in state s_1 and computing function f), the machine will always output o_1 given the

input i_1. In fact, a second machine identical to the first (one that is in state s_1 and is computing *f*) will also always output o_1 given the input i_1.

If cognitive science is correct in characterizing cognition as computation, then human beings are, in fact, biological computers, but, of course, they are considerably more complicated machines than even the most powerful computers built to date. In fact, even if the claim that cognition is essentially computation is true, that does not rule out the possibility that the brain is a quantum computer (though it is not clear to me what that would entail). David Deutsch, however, does not think this is the case. McDermott also views the brain as a classical computer. When citing David Deutsch below, I am not suggesting that McDermott's view is in any way similar to Deutsch's (because I do not think it is). However, I think it is interesting to note that even some quantum physicists believe the brain to be a classical computer (though, I admit, Deutsch is a very peculiar example in that his views are quite different from those of many other quantum physicists). Deutsch writes:

> It is often suggested that the brain may be a quantum computer, and that its intuitions, consciousness and problem-solving abilities might depend on quantum computations. This *could* be so, but I know of no evidence and no convincing argument that it is so. My bet is that the brain, considered as a computer, is a classical one. (Deutsch 1997, 238)[6]

---

[6] It is important to note that Deutsch's interpretation of quantum mechanics follows the many-worlds interpretation, which stresses the reality of possible worlds. That is, on Deutsch's view, every possible choice, every alternative possibility, is, in fact, an actuality in a nearby possible world. In other words, every possibility is actual, but relative to my point of view, it appears as something I could have done, but did not actually do (but to a copy of me in a nearby possible world, what I experience as the actual state of affairs is merely a possibility that could have occurred, but did not). On Deutsch's view, the notion that S_1 (an agent in one possible world) could have done otherwise means that S_2 (S_1's copy in another, nearby possible world) actually did otherwise. S_1's and S_2's choices, however, are still determined in their respective worlds. That is, both S_1 and S_2 are governed by determined, non-probabilistic scientific laws. An agent, on Deutsch's view, is free insofar as her copy in a nearby possible world did otherwise. In other words, a free agent has an open future in virtue of the branching that occurs, but each branch is determined by antecedent events. To the best of my understanding, quantum indeterminacy (the probability-driven branching), on this view, occurs at the level of the entire multiverse while individual worlds are deterministic, law-governed systems. I shall not concern myself with Deutsch's many-worlds interpretation of quantum mechanics (or any other non-collapse

Even if Deutsch's intuition is correct and the brain turns out to be a classical computational system, it does not mean that the brain exhibits von Neumann architecture.[7] The brain may very well turn out to resemble a connectionist machine and thus prove to be a Parallel Distributed Processor (PDP) where neural structures in the brain are responsible for the relevant computations.[8] The details, though perhaps of significance to philosophy of mind and the mind sciences, are not relevant to my discussion. Moreover, the details remain to be worked out empirically as there are many unanswered questions awaiting a proper scientific treatment. McDermott, as already suggested in a footnote, views connectionism as merely one possible architecture for a computational system. He does not accept the view that a connectionist system is non-symbolic, but rather, argues that the symbols of such a system, instead of being ones and zeros, might consist in the firing rates of neurons or something of this nature. For the purpose of this essay, I shall assume that McDermott is correct about connectionism.

It seems quite clear, however, that if human beings are, in fact, such complex computational systems, they cannot, at least in practice, be reset to some initial state and re-run over and over again because the computational processes involved in

---

views). When discussing quantum indeterminacy in the final chapter, I shall adopt the collapse or single universe view.

[7] Jack Copeland lists the central features of a computer with a von Neumann architecture:

> "1 A von Neumann machine is a sequential processor: it executes the instructions in its program one after another.
> 2 Each string of symbols is stored at a specific memory location.
> 3 Access to a stored string always proceeds via the numerical address of the string's location.
> 4 There is a single 'seat of control', the central processing unit or CPU" (Copeland 1993, 192).

[8] McDermott, whose views regarding free will I propose to evaluate, argues that even if the human brain is a Parallel Distributed Processor, this does not entail that digital computers cannot be made conscious. He argues that most connectionist networks are, in fact, simulated on digital machines exhibiting von Neumann architecture. Moreover, a PDP network, according to McDermott, also manipulates symbols and thus is not a good example of a non-symbolic system.

human cognition compute much more complicated inputs, which result in multiple outputs.

Also, as some have argued,[9] the computational processes responsible for cognition may not be confined to the head (the brain), but extend beyond the skull and incorporate various environmental factors in these cognitive processes. Thus, 'resetting' such an embedded system would require the resetting of the environmental circumstances that are part of the system's cognitive repertoire.[10] The system's recorded history (i.e. memory) would also need to be re-formatted and returned to some original setting. Drew McDermott writes: "there is no such thing as an isolated CPU in the real world. In practice, a digital computer is never in the same state twice" (McDermott 2001, 192). However, though impossible in practice, such rewinding would, in principle result in the recurrence (or duplication) of the system's complex web of outputs or behaviours (although I am not certain how a quantum

---

[9] See, among others, the following sources:

Clark, A. (2005). "Intrinsic Content, Active Memory and the Extended Mind." *Analysis* 65, (January). 1-11.

Clark, A. (2003). *Natural-Born Cyborgs: Minds, Technologies, and the Future of Human Intelligence*. New York, NY: Oxford University Press.

Clark, A., Chalmers, D. (1998). "The Extended Mind." *Analysis* 58. 7-19.

Kosslyn, S. M. "On the Evolution of Human Motivation: The Role of Social Prosthetic Systems." *Evolutionary Cognitive Neuroscience*. Ed. S. Platek. Cambridge, MA: MIT Press, in press. 1-24.

Sterelny, K. "Externalism, Epistemic Artefacts and The Extended Mind." *The Externalist Challenge. New Studies on Cognition and Intentionality*. To appear in (ed.) Richard Schantz. de Gruyter, Berlin & New York. 1-25.

Wilson, R. A. (2005). "Meaning-Making and the Mind of the Externalist." *The Extended Mind*. Ed. Richard Menary. Ashgate Press, in press. 1-18.

Wilson, R. A. (2004). *Boundaries of the Mind: The Individual in the Fragile Sciences*. New York, NY: Cambridge University Press.

Wilson, R. A., Clark, A. "How to situate Cognition: Letting Nature Take its Course." *The Cambridge Handbook of Situated Cognition*. Ed. Murat Aydede and Philip Robbins, in development. 1-34.

[10] This is because, as far as I understand this view, the claim is that certain mental processes are not actually performed in the head, but occur outside (in the world). And thus, if we only reset the mental states of the organism (such as the sensory input from the environment), we will inevitably miss those computations that occur outside.

computer would respond to such resetting, other than that it would most likely behave probabilistically).

I shall, however, follow Drew McDermott in his assumption that if human cognition is computation, it is classical in the sense that even though quantum effects are ubiquitous, they occur on the micro-scale and do not affect the computations significantly enough to be considered at the level of cognition.[11] Although all physical systems are fundamentally quantum mechanical, classical computational systems are close enough approximations and thus do not require that quantum effects be taken into consideration. If this sketch of the causal and mechanical nature of computation is correct, then given initial settings and inputs (and assuming complete understanding and knowledge of the program the computer is executing, etc.), the outputs are, in principle, predictable. In other words, the outputs are determined by the inputs in conjunction with the mechanical processes employed in the computation. So, given i_1, there is no sense in which the output o_1 could have been different (i.e. there is no sense in which given i_1, a system in an initial state s_1 computing f would result in o_2 or o_3, etc.). To personify such a system, the system *could not have done otherwise* than output o_1, given i_1 if computing f.

---

[11] Having said this, I will abandon this strategy in chapter four where I shall explore the possibility that quantum indeterminism is in fact a salient feature of a decision-making system. The reason why I assume that the mind is a classical system (or at least a good enough approximation of a classical system) is that the purpose of chapters one through three is to develop a notion of responsibility that is compatible with McDermott's view and so, I shall attempt to stay as true to McDermott's world-view as possible and diverging only when I do not feel such a divergence would in any way be contradictory with what I interpret McDermott to be saying. And again, having said this, I point out that part of this project is indeed an interpretive one and thus, I shall interpret McDermott's view of free will in a manner that McDermott probably did not intend. However, I feel that the interpretation of free will I give is not harmful to the rest of his view and, in fact, in my opinion, McDermott cannot make the claim he does regarding free will (that free will is a self-fulfilling illusion), and hence I provide my interpretation.

It must be kept in mind, however, that such simplified examples cannot capture the complexity of intelligent systems such as human beings, but if human beings are ultimately computational entities, then a complex intelligent system will be composed of numerous interrelated and causally linked subsystems, which, at the fundamental level of computation, may very well remain determined entities.

It may also be the case that the world is not deterministic and that indeterminism reigns at the micro-level, as some interpretations of quantum mechanics seem to suggest. I shall consider such a possibility in chapter four. Presently, I do not think that quantum indeterminacy is pertinent to the following discussion.[12]

Computationalists face other problems as well. They face the incredibly difficult task of explaining the elusive phenomenon of consciousness, the disheartening chore of accounting for qualia, and the daunting challenge of explaining intentionality. These problems are as difficult as they are interesting and although the computational theories of mind cannot, I think, avoid the discussions that have and will continue to arise from these problems, I wish only to explore the problem of free will and responsibility in this essay. To this end, I will assume that solutions to the other problems are within the computationalist's reach (even though this is certainly debatable). The issue of free will and responsibility is an enormous field of inquiry in itself, and although the problems of phenomenal consciousness and intentionality are not irrelevant to contemplations about free will, I am compelled, for reasons of available space as well as simplicity, to explore the problem of responsibility

[12] It is not pertinent because McDermott, as far as I know, does not engage in the disputes over whether the brain is a quantum computer or whether quantum indeterminacy is compatible with free will.

somewhat in abstraction. And although I must describe the computational view of consciousness and the self, I shall not engage in the debates on these very important issues since these merit entire papers of their own.

As a means of sketching the computationalist view of mind, I will briefly outline the self-model theory of consciousness and its implications for the issue of free will and responsibility. The notion of a self-model is, to my mind, a computational version of what in the free will literature is referred to as the agent. In essence, the self-model is that which gives rise to our belief in self-determination.

1.3
What is a Self-Model?[13]

As will become evident from the following sketch, a self-model is a set of computational processes. As theorists who adhere to the self-model view explain, the self-model is responsible for phenomenal consciousness, the experience of a self, and the introspective belief that the mind/self enjoys ontological independence from the body.

Before attempting a fuller characterization of the self-model, it may prove helpful to first consider an explanation of what constitutes a model. Marvin Minsky gives the following definition of a 'model':

> We use the term "model" in the following sense: To an observer **B**, an object **A*** is a model of an object **A** to the extent that **B** can use **A*** to answer questions that interest him about **A**...[Thus] [i]f **A** is the world...**A*** is a good model of **A**, in **B**'s view, to the extent that **A***'s answers agree with those of **A**'s, on the whole, with respect to the questions important to **B**. (Minsky 1965, 426)[14]

---

[13] In the following sections, I shall outline Thomas Metzinger's notion of a self-model in order to set the stage for McDermott's view. However, Metzinger and McDermott do not agree on everything and thus, my main and only focus when considering Metzinger's view will be the notion of self-modelling.

[14] This definition of a model seems to be compatible with both an observer-relative as well as a stance-dependent conception of modelling.

Weather forecasts, for example, depend on the meteorologist's interpretation of various simulations and models. The meteorologist (the observer **B**) makes use of the computational processes responsible for meteorological predictions (physical symbols manipulated by various physical processes inside a machine—object(s) **A***). These computational processes (models and simulations of weather patterns) give the meteorologist (**B**) information about the weather (actual atmospheric patterns **A**).

Minsky goes on to explain that if we have a model of the world **W***, then **M-W*** is the system **M**, which contains the model **W***.[15] Now, it is also possible to have models of models. That is, if **W*** is a model of the world and the world **W**, which **W*** is modelling, actually contains **M**, then **W*** contains a model **M*** of **M**. And **M*** can contain a model **W**** of **W***, and we can always go one step further where **W**** contains a model **M**** of **M***. I would imagine that this modelling of models could, in principle, go on ad infinitum.[16]

Minsky goes on to argue that **M**'s model of himself is "bipartite, one part concerning his body as a physical object and the other accounting for his social and psychological experience" (Minsky 1965, 427). Thus, when we observe something (an object) that is in motion, we either attribute a simple physical force as the source of its movement or a purposeful "self-caused" motion, but rarely both. And so, Minsky claims, our dualistic intuitions stem from this bipartite appearance of our self-models. "*His* [**M**'s] *statement (his belief) that he has a mind as well as a body is the*

---

[15] Minsky uses **M** to denote a man and **M-W*** to denote an internal mechanism **W*** inside of **M**.

[16] I am certain, however, that there are very real physical constraints on how many models can be modelled. Eventually, the computational resources required for such modelling of models may run out making such modelling intractable in practice, but in principle, such recursive introspection can be repeated ad infinitum. Minsky writes: "With interpretative operation ability, a program can use itself as its own model, and this can be repeated recursively to as many levels as desired, until the memory records of the state of the process get out of hand" (Minsky 1965, 430).

conventional way to express the roughly bipartite appearance of his model of himself" (Minsky 1965, 428).

Moreover, even if we have a unified theory of both mechanical and psychological phenomena, Minsky maintains, we will still hold on to these "illusory," bipartite experiences we possess. "[F]or practical, heuristic reasons, [our personal world models] would still retain their form of quasi-separate parts...[because] [t]he primitive notions of physics, or even of neuro-physiology, will be far too remote to be useful in accounting directly for the mental events of everyday life" (Minsky 1965, 428-429).

Next, we are faced with the question of how a self-modelling entity can become conscious of itself. Ray Jackendoff, in "The Computational Mind" (chapter 2 of his book *Consciousness and the Computational Mind*), outlines how a distinction between primary and reflective awareness may emerge. He cites Putnam's point about how the computer analogy helps explain how one can know something without being aware that one knows it. Jackendoff explains:

> Suppose that in order to know some fact *F*, a machine must contain some configuration of computational states *C*. Then, for the machine to be *aware* of knowing *F*, it must be *aware of being in configuration C*. But that requires the realization of some further computational state *C'* that checks whether *C* is present. If *C* is present without *C'*, the machine knows *F* but is not aware of knowing *F*. (Jackendoff 1987, 16)

I understand Jackendoff as claiming that if *C* is present without *C'*,[17] then the machine has primary awareness of *F*, but when *C'* checks whether *C* is present, the machine has reflective awareness of *F* (in other words, the machine has awareness of its awareness of *F*). That is, if one adopts a higher-order theory of thought (as McDermott does) creatures without C' neither exhibit phenomenal consciousness nor

[17] That is, if a fact *F* is represented by or encoded in some configuration of computational states *C* where *C* is not monitored by some further computational state or states *C'*.

can they be self-conscious. Thus, I understand the above as implying that a self-model is a set of processes akin to *C'* (a set of processes that monitor a system's interaction with the world). "That is, the computational device's self-monitoring is a set of processes beyond those responsible for ordinary interaction with the world" (Jackendoff 1987, 16).

1.3.1
Virtual Selves

Having reviewed Minsky's definition of a model and some characteristics of self-modelling, there remain several other features of self-models worth exploring. Thomas Metzinger claims that there are no such entities as selves, but only self-models. "The self-model is an episodically active representational entity, the content of which is formed solely by properties of the system itself" (Metzinger 2000, 289). In human beings, the self-model possesses a neurobiological description (a complex neural activation pattern). The self-model can also be viewed on a more abstract level where the same pattern of neurobiological activity can be described as a complex functional state. Thus, one can take the classical cognitivist perspective when analyzing the self-model. "[T]he self-model could be described as a transient computational module, episodically activated by the system in order to regulate its interaction with the environment" (Metzinger 2000, 290).

According to Metzinger, self-models are not necessarily true. The self-model in human beings is a *virtual* model, which is highly context-dependent. The content of the self-model is simply the best hypothesis about the current state of the system given all epistemic constraints and whatever information is available to the system at

the time. Thus, the content of the self-model does not reflect reality, but amounts to a possibility. And "this possibility is depicted *as* a reality…[t]he actuality of situated self-awareness is a *virtual* form of actuality" (Metzinger 2000, 290).

I understand Metzinger's point about the *virtual* nature of the self-model as simply stating that it is possible that the content of the self-model is not perfectly correlated with the actual state of the external world. That is, for starters, Metzinger does not think that selves actually exist; selves are not something above and beyond the processes involved in self-modelling. In other words, the self-reflexive nature of self-modelling gives rise to the concept of a self. When a system models itself as the very entity doing the modelling, such a system must ascribe to itself a symbol denoting itself in order to carry out the necessary self-reflexive computations. It imputes a self that is responsible for the self-reflexive computations in order to be better able to track such computations. One way (though perhaps not the best way) of understanding this is to imagine that the system adopts what Daniel Dennett calls the Intentional Stance toward itself.[18] If system *S* needs to know what *S* might do in circumstance *C*, it must have a way to predict the possible behaviour of *S*. Adopting the Intentional Stance toward itself allows *S* to make the appropriate predictions and act accordingly. The self, then, on such a view is a convenient way for *S* to track the behaviour of *S*. When Metzinger states that there are no such things as selves, I take

[18] Dennett explains that the intentional stance is a stance we adopt in order to be better able to predict a system's behaviour (since sometimes we are unable to predict the behaviour of a system merely from either the physical or the design stance). One reason for adopting the intentional stance toward oneself might be that it is impossible for a system with a transparent self-model to adopt the physical or design stances toward itself. One difference between the self-referential intentional stance and the stance adopted by an outside observer toward a system, then, is that whereas an outside observer only assumes that the system she is trying to predict and understand has intentional states, but is, in principle, capable of adopting the design or the physical stance toward the system, a self-referential system is incapable of adopting the design or physical stance toward itself because the self-referential system is incapable of modelling itself as causally determined.

him to mean that there is no entity (like a soul, for instance) in addition to the system; the self is simply a system's model of the system represented in its self-model.

If the content of the self-model did not correlate with the current state of the system and external reality approximately enough, the entity with such a "defective" self-model would not be capable of performing many (if any) of the tasks it needs to perform in order to survive. And so, even though the content of the self-model does not perfectly reflect reality, it must approximate it closely enough most of the time. Thus, if system *S* has beliefs about a certain state (or states) of affairs *A*, these beliefs (assuming they are true and not the result of a mistake or misrepresentation of some sort) need to be accurate enough, and if they are not, so much the worse for *S*. For instance, if *S* models itself as being thirsty and believes that the glass containing a clear fluid in front of *S* is water and is sufficient to satisfy the pressing desire for H_2O, it better be the case (for *S*'s sake) that the desire is, in fact, for H_2O and that *S*'s beliefs regarding the contents of the glass are true. However, as already stated, the self-model is a *virtual* model. It just happens to be reflective of the way things really are most of the time,[19] though not all of the time (as the following example demonstrates).

An example of the virtual character of the self-model, provided by Metzinger, is an experiment conducted on a patient suffering from phantom pain caused by a "paralysis" of one of his/her missing phantom limbs. The experiment was performed

[19] I would imagine that self-models whose contents correlate with the current state of the system and external reality approximately enough most of the time would be selected for by evolutionary processes.

by Ramachandran and colleagues who constructed a "virtual reality box"[20] by placing a vertical mirror inside a cardboard box with the top of the box removed and two holes in the front for the patient to insert his/her real and his/her phantom arm. The mirror reflecting the real arm creates the illusion of a second arm. The patient upon being asked to move both arms in unison and once he/she observed that both moved, "felt" the motion of his/her otherwise paralysed phantom arm and thus experienced a relief from the pain caused by the phantom paralysis. When asked to close his/her eyes, however, the phantom limb once again stiffened into paralysis and once more became the cause of discomfort and pain. Metzinger concludes: "What is moving in this experiment is the phenomenal self-model. The sudden occurrence of kinaesthetic qualia in the degraded sub-space of the self-model was made possible by installing a second and perfectly superimposed source of 'virtual information,' restoring, as it were, the visual mode of self-representation and thereby making this information volitionally available" (Metzinger 2000, 291).

1.3.2
The Phenomenal Experience of Self ("Mine," "Me," and "Myself")

According to Metzinger, all representational states embedded in an active phenomenal self-model gain the additional higher-order property of phenomenal "mineness" (a pre-reflexive, non-conceptual sense of ownership). Thus, a self-model enables a system to represent itself to itself as an agent. Such self-representation allows the system to differentiate between its own and foreign actions.

[20] See Appendix 1 for an image and further explanation.

Phenomenal selfhood is a fundamental form of non-conceptual self-knowledge (an inner acquaintance) that precedes any higher forms of cognitive self-consciousness.[21]

The third phenomenal target property Metzinger mentions is perspectivalness, which is "the existence of a single, coherent, and temporally stable model of reality that is representationally centered on a single, coherent, and temporally extended phenomenal subject" (Metzinger 2000, 296), where a phenomenal subject is a model of the system *as experiencing*. Perspectivalness, then, is the centeredness of a subject. Thus, any system that is in possession of a self-model can become the object of its own attention, concept formation, and self-directed actions. Such a system is capable of modelling the differentiation between the processes representing environment-related and system (itself)-related information. Hence, operating under a model of reality organized in a perspectival fashion enables an information-processing system to generate an entirely new class of actions, actions directed toward itself.

1.3.3 Transparency

It may prove useful to clarify Metzinger's usage of the term 'transparency.' According to Andrew Brook and Paul Raymont, the transparency thesis is "the claim that we are not directly conscious of our own experiencings…Tye, for example, says that when we hear something, we cannot be conscious of the auditory experience, just what it represents" (Brook & Raymont 2006). William Lycan writes: "Harman

[21] Although it appears as though Metzinger's analysis at this point is incompatible with McDermott's higher order theory of consciousness, I shall ignore this difference as it does not detract significantly from my present purpose, which is to spell out some of the general features of self-modelling.

(1990) offers the transparency argument: We normally 'see right through' perceptual states to external objects and do not even notice that we are *in* perceptual states; the properties we are aware of in perception are attributed to the objects perceived" (Lycan 2006).

> The self-model's representational structure is transparent to the self-model.
>
> For this reason we are permanently operating under the conditions of a "naive-realistic self-misunderstanding": We experience ourselves as constantly being in direct and immediate epistemic contact with ourselves. What we have in the past simply called a "self" is not a non-physical individual, but only the content of an ongoing, dynamical process—the process of transparent self-modelling…The phenomenal self is a virtual agent perceiving virtual objects in a virtual world. (Metzinger 2000, 299-300)

So, the conscious self is an illusion (the notion of the self as illusion can perhaps be better understood as a virtual self arising due to the representational content of the self-model), but it is an illusion that belongs to the system as a whole and thus not to any individual self (because there is no self aside from self-modelling processes), which makes it "an illusion that is *no one's* illusion" (Metzinger 2000, 301).

An interesting comparison (one that may shed some light on the kind of illusion we are dealing with here) can be drawn between the very strange illusion of self and the similarly perplexing "moon illusion" or "moon effect" where the moon seems larger when it is near the horizon than when it is high in the sky.[22] As far as I know, everyone experiences this effect. Many different explanations have been offered for this optical illusion (the veracity or falsity of these explanations is irrelevant for my purposes), but the one thing everyone agrees on is that the moon does not actually get larger (and does not, at any time, draw close enough to the earth to account for the effect) when it is spotted near the horizon. (See appendix 2 for another example, the Ponzo Illusion). The interesting thing about such illusions is

[22] I would like to thank Professor Wesley Cooper for suggesting this vivid comparison.

that even if we know them to be just that, namely illusions, we cannot help but experience them. Similarly, the self, on the self-model view, can be seen as such an illusion. We persist in the illusory experience of being selves whether we know the experience to be illusory or not (we all have this illusion, but we cannot shake it).[23]

Although opaque states do exist, transparency is one of the (if not *the*) most important constraints in that it provides us with a theoretical understanding of what phenomenal experience really is. Even though complex neural patterns are ultimately responsible for consciousness, the conscious "self" is not aware of these patterns qua patterns nor is it aware of the existence of neurons (the only reason I believe that there are neurons in my head is because I have been told this fact, but most of us never actually go to all the trouble necessary to check for ourselves). "This medium is transparent insofar as the subpersonal processing mechanisms contributing to its current active content are attentionally unavailable to high-level introspective processing" (Metzinger 2003, 169).

For example, phenomenal colour vision is transparent because we are unable to direct our attention to the ongoing activity of the relevant processing mechanisms in our visual cortex. There cannot exist a conscious self-representation that is solely characterized by opaque content; cognitive self-reference, according to Metzinger, must take place against the background of transparent, pre-conceptual self-modelling. On the level of phenomenal representation, a transparent phenomenal self-representation is characterized by the fact that the transparent self-model is unable to

[23] The Illusion of free will is also such an illusion, one we cannot shake when we are actively engaged with the world.

discover the difference between self-representational content and self-representational vehicle. In other words, the self-model is not recognized as a model by the system.

> For intended phenomenal self-simulations we always find two transparent phenomenal properties, namely, mineness and agency, which lead to an overall state in which the content of these states is clearly marked out as my *own* mental activity, which has been deliberately created by *myself*. Thoughts about the future of this organism now unequivocally become *my own* thoughts about *my own* future. (Metzinger 2003, 342)

The self-model view, then, becomes, not only a powerful explanation of self and agency, but also a convincing account of why we continue to hold on to the Cartesian intuition that minds are non-extended substances.

The transparency constraint gives rise to a naïve realist view of ourselves. Parts of our bodies (the neural correlates giving rise to the phenomenal experience of our own embodiment) act as an object emulator, emulating the phenomenal body (the body we feel ourselves inhabiting). The transparent nature of our self-models allows us to take the spatial character of bodily experience for granted, "as if it were not a representational construct but something to which we had direct and immediate epistemic access. The same mistake is then made with regard to phenomenal cognition, the internal representation of certain cognitive processes on the level of conscious self-simulation" (Metzinger 2003, 381). The "thinking self" arises out of the necessity of generating a mental model of ourselves as beings producing thoughts and conceptual knowledge.[24] This "thinking self" introduces a fundamental chasm in the conscious self.

> What makes [the "thinking self"] a highly successful new virtual organ is the fact that it possesses an entirely different function *for* the system than the bodily model of the self: It has to make those cognitive processes that need to be constantly monitored available for self-directed attention and higher-order cognition...this partition of the unconscious self-

[24] This happens because "[w]e are systems which have to explain to themselves how it was possible that we can carry out abstract, cognitive operations using non-sensory, second-order simulata" (Metzinger 2003, 381).

> model...cannot directly be connected with the phenomenal image of our body on the level of conscious experience and therefore the organism cannot *own* it. (Metzinger 2003, 381)

The self, according to Metzinger, is not an entity with any real ontological status. Metzinger explains:

> Under a general principle of ontological parsimony it is not necessary (or rational) to assume the existence of selves, because as theoretical entities they fulfill no indispensable explanatory function. What exists are information-processing systems engaged in the transparent process of phenomenal self-modeling. All that can be explained by the phenomenological notion of a "self" can also be explained using the representationalist notion of a transparent self-*model*. (Metzinger 2003, 337)

Even though the transparent nature of the self-model gives rise to the Cartesian intuition that minds are non-extended substances, what we naively call the self, is really nothing more than the content of our self-models.[25]

To summarize, then, the conscious self, on the self-model view, is an illusion (a virtual construct), which paradoxically is no one's illusion since there are no selves; all that exists is the self-model. A self-model consists of a set of processes that monitor a system's interaction with the world and it could be described as a transient computational module, episodically activated by the system in order to regulate its interaction with the environment. Self-models are not necessarily true. In fact, the self-model in human beings is a *virtual* model, which is highly context-dependent. The self-model provides us with the feeling of centeredness and a feeling of the continuity of an embodied self (by giving rise to the phenomenal properties of "mineness," "selfhood," and "perspectivalness"). Also, the self-model's representational structure is transparent to itself. And, the self-model is responsible

[25] In a sense, the self is a real entity in the world insofar as it can be objectively picked out from a third person perspective. That is, the processes responsible for the self must have actual neurological correlates. However, there is no such thing as a soul or a Cartesian non-extended immaterial res cogitans on the self-model view.

for one's "introspective" intuition that one's mind is a separate entity from one's body.

Chapter Two
McDermott on Self-Modelling, Qualia, and Free Will

2.1
Self-Modelling and Deliberation

McDermott sets out his explanation of the self-model by making two crucial points: first, that every belief one has about oneself derives from the self-model and second, that the beliefs derived from the self-model, including the belief in a self, do not have to be true in order to be useful. Furthermore, the self's existence is fully dependent on the self-model. That is, the self does not exist prior to being modelled (McDermott 2001, 4-5). Thus, McDermott believes that "introspective intuitions about internal representation are unreliable" (McDermott 2001, 88). Moreover, he claims that the unity of the self—that is, the introspective intuition that "the mind is a meeting place for representations from all modalities expressed in a single internal 'conceptual structure'" (McDermott 2001, 88)—is an illusion.

McDermott examines the notion of the self-model with the aid of a thought-experiment about an intelligent robot and its ability to deliberate about its own choices. The robot (lets call it Rosie) must obviously have a way to model the world (its environment) in order to make choices. The difference between what Rosie does and what computers normally do is that Rosie models the situation she is currently in and so, Rosie models her undertakings on the assumption that *Rosie* will be carrying out those actions. In order to be able to make such an inference, however, Rosie's model will need to include, among other various symbols, a symbol (McDermott uses 'R') denoting "herself." In other words, the robot models itself along with the various other models the robot is running such as, for instance, the robot's world model.

In McDermott's example, the robot finds itself in a situation where it is standing next to a bomb with a lit fuse. The robot's model of the world is accurate enough to predict that if R is standing next to B (the bomb with a lit fuse), then, if R maintains its current position relative to B, R will be destroyed. Assuming, as McDermott assumes, that R is currently running a "self-preservation" module, which is a standing order to avoid damage, R computes that unless R moves away from B, R will be destroyed. And so, R rolls away from the bomb.

The deliberative process described above is a purely mechanized process, meaning that given the environmental circumstances modelled by R (that R is standing in the vicinity of a bomb with a lit fuse) and given R's instructions (to avoid damage), R will "choose" to execute a series of actions leading to R's eventual rolling away from the bomb in an attempt to obey its "self-preservation" program. R's actions, then, are entirely caused by events. In reality, *R could not have done otherwise than R actually did.* This is precisely how I shall be interpreting McDermott's vision of free will and this is, in part, the reason why I will interpret McDermott as an illusionist about free will. The sequence laid out above "is a straightforward causal chain, from perception, to tentative prediction, to action revision" (McDermott 2001, 97).

McDermott argues that the causal chain (from perception, to tentative prediction, to action revision) cannot be represented accurately in R's self-model because the making of tentative predictions involves the model itself.

> The model could not capture this causal chain because then it would have to include a complete model of itself, which is incoherent. In other words, some of the causal antecedents of R's behavior *are situated in the very causal-analysis box* that is trying to analyze them. The robot might believe that R is a robot, and hence that a good way to predict R's behavior is to simulate it on a faster CPU, but this strategy will be in vain, because this particular robot is

> itself. No matter how fast it simulates R, at some point it will reach the point where R looks for a faster CPU, and it won't be able to do that simulation fast enough. (McDermott 2001, 97)

So, the strongest conclusion the robot is capable of reaching is that "'If R doesn't roll away, it will be destroyed; if it does roll away, it won't be.' And then of course this conclusion causes the robot to roll away" (McDermott 2001, 98). The robot must model itself in a different way from other objects. Whereas the behaviour of other objects is modelled as caused, R's behaviour must be modelled as open or as still being solved for. Hence, the self-model is responsible for the robot's "conviction" that it is exempt from causality. McDermott's definition of free will, then, boils down to the following: "A system has free will if and only if it makes decisions based on causal models in which the symbols denoting itself are marked as exempt from causality" (McDermott 2001, 98). Free will, on this view, is an illusion caused by self-modelling processes. McDermott aims to show that this illusion is a self-fulfilling illusion, meaning that once a system models itself as being free, it becomes free in virtue of the belief in its own freedom.[26]

2.2
Qualia

McDermott's vision of free will is intimately connected with his adoption of the self-model view. His account of free will, as I interpret it, runs parallel to his explanation of qualia. I do not think that on his view, phenomenal consciousness is necessary for free will, however. But, free will, as I understand McDermott's view,

[26] I do not think that the illusion of freedom is a self-fulfilling one. In fact, I do not think McDermott successfully shows that it is. I shall argue that what McDermott does succeed in showing is that the experience of freedom is real.

being a product of the self-model, arises under similar conditions to those that give rise to qualia.

McDermott does not treat qualia as mysterious non-physical mental entities. Rather, since, on his view, qualia come embedded in a framework of comparisons with other qualia, the comparison mechanism responsible for picking out one quale from another, based on the quale's content, is the result of the system's inability to make further discriminations in tracking a given quale (i.e. a red quale). That is, a red ball, a fire truck, and a red wall will be labelled as indistinguishable in colour by such a system.

My understanding of McDermott's take on the free will problem is that the feeling of free will, like the quale of red, is something that exists only in the self-model (or only because a system has a self-model). That is, I am compelled to view McDermott's notion of free will as the existence of an experience of freedom (but nothing more than that), which accompanies certain types of actions.

Qualia, on McDermott's view, though "real," exist only in self-models. That is, because there is no real need for qualia in an ordinary computational system, qualia are brought into being only by the self-reflexive process of self-modeling. An intelligent robot, equipped with a variety of sensors, might experience the functional equivalent of some quale (i.e. a certain kind of pain, for example).

McDermott gives the example of a robot (R) designed to carry humans out of burning buildings. Such a machine, being expensive, will need to be equipped with heat sensors that will alert it when the temperature it is being exposed to gets too high. In addition, as part of the overall program, the robot, in McDermott's example,

has two distinct goals: the first goal (G1) is a self-preservation module while the second goal (G2) is the instruction to search for and ensure the safety of humans trapped in the burning building.

As R searches for humans, R's heat-sensors record ever-increasing temperatures. Given that R has not detected humans in the building and given that it computes the probability of finding humans to be very low, it may opt to abandon G2 and focus ever more increasingly on the instructions contained in G1. Once the robot begins to follow instructions of G1, and when the temperature sensors report "extreme heat," a goal may be set up to flee, even though it may not be acted on. Even if R continues its search for humans, the sensor report is impossible to ignore; the sensor signals demand computational resources to evaluate whether it is necessary to act on them. As long as R decides to stay in the flaming building, the signals carrying the relevant information about the surrounding heat and the state of R's robotic body will be labelled as "unpleasant but bearable." "At this point we can conclude that the robot's perception of the fire has something like a quale of unpleasantness…I mean that however the state is represented, it is classified as 'to be avoided or fled from,' and it is so classified *intrinsically*" (McDermott 2001, 102).

What McDermott means by *intrinsically*, can be understood with the aid of the following example: R (still in the burning building) encounters another robot (C). C is the intelligent controller of the building and now that the building is in flames, there is no reason for C to continue its existence and thus C calmly awaits its own destruction. If C were to ask R why R is moving so carefully toward the exit (assuming that R has now fully adopted G1), R may respond that it is exiting the

building. If asked why R is exiting the building, R might respond that if R remains in the building, the extreme heat will be the cause of R's destruction and thus, R is carefully moving toward the exit in order to avoid such massive damage. If C is persistent and curious about R's behaviour, C might ask why R wants to avoid destruction (that is, its own death). At that point, R may not have a reasonable answer to give. Its program may include a ready answer along the lines of "I was designed to avoid damage and my own demise," but such a response would only be the cause of its behaviour and not a reason for it. Thus, when confronted with the question, R's reasons will run out. Hence, the avoidance of heat in order to avoid damage is so classified *intrinsically*. That is, the avoidance of damage, as far as R is concerned, is a *good* in itself, which does not require any further reasons.

R's goals, McDermott explains, will be labelled as preferences. That is, R's goal gets labelled as a preference by being assigned a certain probability to a course of action or some behavioural routine. If a creature (like R) is presented with conflicting goals (such as G1 and G2), it must have tags or labels assigning relative values to various situations. Given that R abandons G2 (because of the low probability of finding any more humans), the relative value of G1 rises. There is no point in having these values questioned by R. "A creature that could really question the value of everything would never act" (McDermott 2001, 103).

McDermott anticipates the objection that although R may manipulate symbols that label the input from its heat sensors as something *like* "unpleasant and to be avoided," R does not exhibit anything like an "experience" of heat or pain. He responds by stating that every quale comes embedded in a framework of comparisons

with other qualia. Such comparison mechanisms are not accessible to consciousness, but manifest themselves in the conscious sensations that the brain makes use of. This prevents the brain from trying to think about how it works. According to McDermott, there is a great amount of content in a pain sensation (i.e. "This pain X is like another pain Y," "This pain X is unpleasant," "This pain X tends to get better if I don't sit down," etc.). The actual sensation, then, is just a way of labelling that content. The comparison mechanism plays a primary role in the system; the sensations are merely an aspect of the introspection about the comparison mechanism.[27]

However, it is impossible to compare one person's quale with that of another (i.e. S_1's red quale with S_2's red quale) because what red looks like to S_1 is determined by S_1's self-model while what red looks like to S_2 depends on S_2's self-model. McDermott does not appear to grant qualia and free will full ontological status. McDermott makes use of a telling analogy: he imagines someone (call her Sue) contemplating her ceiling fan and wondering whether the blades are oriented parallel to the axes of a complex-number plane. If once told that the plane is a mathematical abstraction, which does not actually exist in real space, Sue poses the further question of whether the blades would be parallel to the axes of the complex plane at some point were they to rotate, we should conclude that Sue has not understood. The same can be said, McDermott claims, for inter-subjective qualia

[27] It seems a bit odd (perhaps circular) to explain the quale of unpleasantness by postulating a system that labels certain data-structures by means of qualia. I am not certain that McDermott's explanation of how qualia are represented by a computational system is complete (and I have my doubts that it is convincing). Perhaps it is important to keep in mind that the robot has only "qualia-like" representations. That is, R has "as-if" qualia, but not actual sensations. As mentioned above, the problem of qualia is an issue in need of further debate. For the purpose of this essay, I will assume that there is a more detailed explanation of qualia available to McDermott, one that is compatible with his current approach.

comparisons. Mental entities, according to McDermott, are quasi-fictional. "In important ways, fictions are exactly what qualia are, useful fictions with a grain of truth (because they are attached to real sensory events)" (McDermott 2001, 157). I propose to analyze free will in a similar manner. Freedom of the will, like a quale,[28] is a useful fiction, but a fiction nonetheless. McDermott attempts to give reality to these fictions and perhaps he succeeds, but only to a certain (quite limited) degree. Before continuing our analysis, I am compelled to take a short, but important detour.

2.3
Symbols and Semantics

It will prove useful to skim over McDermott's understanding of symbols and semantics (though I shall limit this discussion only to the aspects that are relevant to McDermott's insight about qualia and their free will analogue).

In response to thinkers like Searle, McDermott defines computation objectively. Searle argues that computers are just systems that people use to compute things and so, without people to interpret the inputs and outputs being computed, the computer is just another physical system. In other words, computers are observer-relative.

McDermott defines a computer as a physical system whose outputs are a function of its inputs. Since inputs and outputs are not usually labelled for our convenience in natural systems (as they are in system diagrams), we must pick out the input with some sort of objective criterion. McDermott suggests that the most obvious way is causation: the input causes the output and not the other way around.

[28] In fact, on my understanding of McDermott's view of freedom, free will amounts to being nothing more than just the phenomenal experience of freedom.

"More precisely, when the system is in a certain input state, that will cause it at a later time to be in a corresponding output state" (McDermott 2001, 169).

A system can be said to compute a function if its states can be interpreted as the inputs and outputs of that function. The system's inputs and outputs can perhaps be interpreted as performing various other functions as well, but McDermott claims that whether they must be interpreted that way is a separate issue and whether they are interpreted by anyone at all is irrelevant to his objective definition of computation.

McDermott also explains his notion of a "bare symbol," a symbol viewed purely syntactically without reference to semantics. A symbol can have multiple occurrences (that is, a symbol type can be expressed by numerous symbol tokens of that type). There must also be a symbol site (either a position in space or some other non-spatial site).[29] McDermott defines a *symbol site* as "a set of mutually exclusive alternative states of a system *at a particular point in time*" (McDermott 2001, 82). Thus, a symbol site can be occupied by different states; the state that actually obtains is called the *occupier* of that site.

What about the symbol token? McDermott gives an example of a Turing machine scanning a tape. The symbols the machine uses are blackened squares that reflect very little light (we can use the digit '1' to stand for such a square) and white squares, which reflect a lot of light (we can use the digit '0' to represent the white squares). McDermott imagines the following Turing machine program:

if X=0 and see 1 then write 1; X ← 0; go right

[29] An example of a symbol site that is not a location in space is the vibration of a string at several frequencies simultaneously: "If we focus on the cases where the amplitude of the vibration at frequency *f* is either below 1 cm or above 2cm, then the set {'below 1 cm,' 'above 2 cm'} would be a symbol site" (McDermott 2001, 182).

if X=0 and see 0 then write 1; X ← 1; go right[30]

'X' stands for the machine's memory, which consists of a single integer (either a 1 or a 0). The instruction for the first line means: if X contains a 0, and 1 is the symbol under the scanner, then replace it with a 1, and store a 0 in X; then move scanner to the right. McDermott continues:

> [S]uppose the Turing machine reads a 1, writes a 1, moves left, and sees another 1. Is it looking at the same symbol? Clearly not. But if it now moves right, our intuition says, it will be looking at the same symbol it saw originally…These examples show how tricky it is to get from symbol sites to actual symbols. (McDermott 2001, 184)

McDermott describes what he refers to as the *precursor relationship* in his attempt to define a symbol token. He writes: "Symbol site S_1 is a *precursor* of symbol site S_2 if S_1 precedes S_2, there is a one-to-one mapping between the sets for S_1 and S_2, and the element of S_2 that occurs is caused by the element of S_1 that occurs" (McDermott 2001, 184). Symbol tokens, then, are symbol sites that are causally linked by the precursor relation.

Although McDermott spends much more time on other technicalities, I wish to move on to his view about semantics. The symbol tokens, though they may not have meanings in the Turing machine example described above, certainly must have meanings in human brains (if it is the case that the computational account of the mind is correct and thus that the mind is a result of computations in the brain).

McDermott provides us with the example of a robot using symbols to denote certain objects in its environment. In McDermott's example, the robot is a guardian of a laboratory and its task is to protect the lab from intruders. Its visual system detects something that it recognizes as a human. The robot (R) labels the human it

[30] This is McDermott's example taken from page 183.

detects as X29 (this is an arbitrary symbol R assigns to the intruder). Based on the information provided by its visual system, R also labels X29 as female, X29's height as being 1.8 m, etc. "'X29' refers to the unknown woman in the laboratory, not because there is a decoding in which that symbol is mapped to her (although there is such a decoding), but because the woman is the cause, in an appropriate sense, of the symbol structures" (McDermott 2001, 197).

An objection to this view, one notably voiced by Searle, is that the symbols only mean something to the humans interpreting them, but that the symbols themselves are meaningless to the robot just like the words and entire sentences contained in a book are meaningless to the book because both the robot and the book lack original intentionality.

McDermott distinguishes informational meaning from intentional meaning. The term *intentionality* is used to refer to the capacity of mental states to "be about" or "be directed at" something. For instance, an astronomy textbook may contain a chapter that is "about" the sun (the book contains certain symbols arranged in given patterns, which can be interpreted as being about or describing the sun). That particular chapter of the book, then, is about the sun. However, the book is said to have *derived intentionality* because "its meaning depends entirely on the interpretation put on it by the civilization that made it" (McDermott 2001, 198). Similarly, when I think about the sun, there are certain physical processes going on in my head, which "are about" something (an object) that is approximately 146 million km away (the notion of intentionality is quite mysterious). *Informational meaning* can be defined as follows: "a physical event *E* means something about an antecedent

event or situation *P* if the occurrence of *E* changes the probability of *P*. That is, it is evidence for or against *P*; it provides *information* in the technical sense about *P*" (McDermott 2001, 198). McDermott argues that a tablet containing a lost, ancient written language might still mean something to the archaeologists studying it[31] in the same way that strata in the earth mean something to geologists. For instance, a layer of iridium in the ground refers to the collision of a meteor with the earth.

According to McDermott, if we imagine a robot that has certain "beliefs" about a person (i.e. a woman) who happens to be within the robot's visual field, we could say that "[t]he robot's data structures refer to the woman the way a layer of iridium in the ground refers to the collision of a meteor with the earth, except that the causal links in the former case are much 'thicker'" (McDermott 2001, 198-199) because the data structures have a considerable effect on the robot's behaviour, which, in turn, is the cause of an intricate sequence of further events. McDermott speculates that it may turn out that intentionality can be explained in terms of informational meaning (thereby erasing its seeming distinctiveness from intentional meaning). However, he continues:

> Let's back away from that claim for now. I think it's correct, but we are not yet in a position to argue for it, simply because cognitive science is still too primitive. For now, we should focus on the technical problems involving informational meaning in robots. (McDermott 2001, 199).

McDermott uses the following thought experiment as a means of considering problems regarding informational meaning in the laboratory-guarding robot: he imagines that sometime in the far future, human beings give up technology and no longer make use of the English language, which, at that time, is lost. The guardian

[31] Archaeologists can infer a great deal about a culture studying its writing methods and implements even if the symbols used in the written language are unknown to them.

robot, still functioning at this faraway time, is being examined by visiting aliens who try to figure out what the symbols the robot is using actually denote (what they actually mean). Of course, they are unable to ask any human for assistance. So, they open the robot's head and notice a triple symbol structure: α-X29-♀[32] where the third symbol is sometimes a ♂. The aliens realize that the middle symbol is always different and that ♀ is always correlated with the robot's detecting a human female while the symbol ♂ occurs when a human male is located. The aliens conclude that α stands for gender, X29 (or any other middle symbol used) denotes a specific individual, and ♀ means female. It also turns out, however, that the symbol ♀ comes up every time the robot sees an alien (that is because, as McDermott assures us, aliens closely resemble human females). Thus, the symbol ♀, under these new circumstances, means either human female or alien.[33] Hence, McDermott concludes that a symbol means whatever the people who started using it think it means and whatever it is most reliably associated with in its environment. McDermott suggests that "exactly what symbols denote is highly context-dependent" (McDermott 2001, 202). He continues:

---

[32] McDermott uses the symbols 'α-X29-β,' but I will go ahead and make use of 'α-X29-♀' and 'α-X29-♂' for clarity.

[33] McDermott's view is quite similar to Daniel Dennett's notion that original intentionality is merely a myth and that everything can at best possess derived intentionality. In "Evolution, Error, and Intentionality," Dennett imagines a vending machine (a "two-bitser"), which takes US quarters. When a quarter is inserted, it goes into state Q. It can "be mistaken" for if an imitation quarter K is inserted, it also goes into state Q and sometimes when a real quarter is inserted, it fails to go into state Q. What makes the device a quarter detector, rather than a slug (K) detector, according to Dennett, is the shared intention of the device's designers, builders, owners, and most users. "It is only in the environment or context of those users and their intentions that we can single out some of the occasions of state Q as 'veridical' and others as 'mistaken'" (Dennett 1989, 291). So, if the device is moved to Panama, the shared intentions of the users will make it into a quarter-balboa acceptor (quarter-balboa's are like US quarters in shape, size, weight, etc., but are legal tender in Panama). The two-bitser, then, is a quarter detector (going into state Q) due to its intended function. If it is sent to Panama to serve the function of being a quarter-balboa detector, then it can be said to go into state QB.

> [I]f the brain is essentially computational, then the events inside it depend on the formal properties of the symbols encoded in its states...A symbol denotes an entity or relationship outside the brain because it occurs inside a symbol system whose most harmonious semantics assigns that meaning to the symbol. A semantics is "harmonious" if it provides a coherent story about the relationships between symbols and sensorimotor events. (McDermott 2001, 202)

2.4
Qualia and Free Will

Having sidetracked slightly into McDermott's notion of symbols and their meaning, we can now return to the issue at hand, namely qualia and their relationship to free will. As already stated, McDermott defines free will as the self-model's modelling of itself as exempt from causality. In *Mind and Mechanism*, McDermott explains qualia in light of his conception of free will and I think that it is best to analyze his view of free will in light of his explanation of qualia.

McDermott claims that in order to have a sensation (i.e. an experience of red), "there must be actual symbol structures in the part of the brain implementing the self-model that actually express those contents, including those similarity relations [produced by the comparison mechanisms responsible for the sensation]" (McDermott 2001, 203).

The self-model, as it occurs in human brains, will be a result of neural patterns. The patterns may depend on whether a neuron is firing, whether a group of neurons is firing, or the frequency at which they fire may all serve as symbol sites, and of course "[t]here may be other symbol sites we haven't guessed the existence of yet" (McDermott 2001, 203). These patterns, whatever they may be, should have sub-patterns denoting the entities referred to (such as "I," "experience," colour shades, shapes, etc.). The fact that a set of symbols (neural patterns) denotes a self-

model depends not only on how it works, but also on how it is connected to the rest of the system.

> [F]inding the individual entities denoted by the symbols in the self-model is not really different from finding more visible entities in the physical world. "This quale" will denote a sensory event because it was caused by that event and the beliefs about it (such as its duration) will be true of that sensory event. Generic terms (such as "experience" or "this shade of orange") will be grounded in much the way the word "female" is, by finding the property or set of objects that the symbol actually tracks. (McDermott 2001, 204)

According to McDermott, what requires explanation when thinking of semantics is the link between symbols and their denotations. For instance, he states that we cannot draw conclusions about the participants of the Six-Day War from an input of data denoting wind velocity, barometric measurements, etc. But we can use these inputs to draw conclusions about the weather because the symbol structures actually denote these weather readings due to the causal relationship between the objects (the actual wind velocity, etc.) and the input data.

McDermott tells us that the beliefs occurring in the self-model are self-fulfilling. For example, "[a] belief in an ordinary pain is part of the causal chain that makes the belief true. An 'erroneous' pain report brings exactly the same chain into existence, with a different first link" (McDermott 2001, 208). Thus, McDermott concludes, the difference between a true and a false sensation report vanishes.

Perhaps the difference between a true and a false pain report is insignificant because, after all, regardless of whether the pain is true or erroneous, the experience of the pain makes it real; the reality of a painful experience does not depend on the appropriate causal chain as much as on the fact that the pain causes me to suffer—that it has a certain phenomenal character—(although usually the appropriate causal chain brings about the corresponding experience). But is this self-fulfilling nature of the content of the self-model true for every mental entity residing in the self-model?

Does the same argument work for the experience of colour, for instance? In a sense, I think it does. After all, whether I hallucinate a vivid green or experience it because I am viewing a green object makes no difference to the fact that the green of my experience is that which I am experiencing. Thus, even if the green quale is a fiction of the self-model, as a quale, it affirms its existence and is thus self-fulfilling. However, there is an enormous difference between seeing a green ball and hallucinating a green ball. The mental entities that come into existence in the self-model are real insofar as they are objects of experience, but not all of these self-fulfilled entities actually exist in the real world. In a sense, of course, since these entities are the result of real symbol structures in the self-model, they exist in the real world as parts of the self-model, but what I wish to argue is that not every mental fiction has the same degree of reality.

Although the difference between a true pain and an erroneous pain does vanish since, for all intents and purposes, a pain is painful and causes suffering regardless of how it came to be, the causal genesis of the pain is not unimportant, it does matter. For instance, an intense pain in my stomach caused by food poisoning, though capable of inflicting the same amount of suffering as an "imagined" or erroneous intense pain in my stomach, is, I think, more meaningful somehow. The stomachache with a real-world origin carries certain information about an event or a state of affairs that might be of importance to the system. That is, a sharp pain caused by a random firing of a cluster of neurons is just *that*, "a sharp pain caused by a random firing of a cluster of neurons." It does not seem to refer to anything else. A sharp pain caused by spoiled food (or by a knife wound), on the other hand, is linked

to an event outside the self-model. Such a pain can, I think, be said to actually denote something other than itself, like damage to the body.

I belabour this difference because I think that it is a real difference and that it is crucial to McDermott's conception of free will. Although in a sense McDermott is correct in claiming that the many fictions of the self-model acquire some reality by the mere fact of being modelled, the self-fulfilling reality of many of these mental entities is comparable to the ontological status enjoyed by fictional characters such as Hamlet. We talk about Hamlet, we quote him, actors imitate (or at least portray) him, we experience emotions toward him, we react (on his behalf) to the treacherousness of Rosencrantz and Guildenstern, etc. In a sense, Hamlet exists. In a sense, mental fictions are real. But there is a difference between Hamlet and a real person like Frederik André Henrik Christian (who really is the prince of Denmark),[34] for instance, as there is a difference between being modelled in the self-model and being represented in the self-model based on input data corresponding/referring to something that is in existence outside and independently of the self-model.

Free will, on my interpretation of McDermott's view, is a fabricated self-conception originating in the self-model. Unlike Hamlet's fictional existence, however, the illusion of free will is a useful deliberative tool in that it allows the system to avoid modelling itself in complete detail because the system, believing itself to be free and thus believing its future to be open and still being solved for, need

---

[34] Frederik André Henrik Christian, the Crown Prince of Denmark, born on May 26, 1968 in Copenhagen, Denmark is the eldest son of Queen Margrethe II and Prince Consort Henrik. Frederik is the heir apparent to the Danish throne.
[Information acquired from: Wikipedia. (February 7, 2007). "Frederik, Crown Prince of Denmark." GNU Free Documentation License.
http://en.wikipedia.org/wiki/Frederik,_Crown_Prince_of_Denmark. (accessed on February 15, 2007)].

not model itself completely. McDermott argues that the system's complete modelling of itself would result in an infinite regress where the system would model itself modelling itself modelling itself, ad infinitum. Such infinitely regressive self-modelling would require infinite resources, to which no finite being has access.

Perhaps another way of understanding the problem of regress is in terms of Karl Popper's argument stating that complete prediction in a classical, deterministic system is impossible if the predictor is part of the system. Popper (1950a) considers a mechanical system *A* and a predictor *B* attempting to predict *A*. *B* can only predict *A* if it can calculate its interference with *A*.[35] That is, *B* must include its act of predicting *A*, and the consequent effect on *A* of such predicting, in its prediction of *A*. One way for *B* to compute its interference with *A* is to study its interfering parts *B′* and their interaction with *A*, but this implies that *B* now needs to study the system *A* + *B′* instead of *A* and the same problem arises again.

The other way for *B* to assess the way in which it interferes with *A* is to study its interfering parts *B′* and their interaction with *A* on the basis of predictions about itself. But this is problematic because a predictor cannot have such self-knowledge about itself because a predictor's knowledge at the time it gives the explanation, in order to explain in detail its own past, must exceed its knowledge at the time for which it gives the explanation. And at any particular instant of time, for any calculator *C*, according to Popper, there will exist a "spacious present" of *C*, which is the minimum amount of time it takes *C* to know what has happened to it. The "spacious present" of *C* divides time for *C* into past and future where *C* can answer

[35] This is because *B* is a part of the system it is trying to predict.

every question asked of it about the past, but cannot provide answers about the future part of the "spacious present." In other words, the "spacious present" is divided "into a closed or fixed past, and an open not fully determinate future. (And this does not only refer to *C*'s knowledge of itself, but also to its knowledge about its 'closer environment')" (Popper 1950b, 193). This, in turn, implies that *C* cannot ever have complete knowledge of itself. That is, it cannot have knowledge of its own state until that state has passed. Therefore, if Popper is correct, McDermott's robot is incapable of modeling itself fully and even if the robot is a determined system, it has no choice but model itself as though it were free.

The experience of freedom (the experience of deliberation between two or more open choices and the consequent selection of one choice over another) amounts to a belief in the openness of one's future and the belief in one's own causal power to choose between competing alternatives. The experience of freedom, which stems from the beliefs in the openness of one's decision-making process and one's capacity to freely choose between several options is similar to the experience of a red thing in that both types of experiences are the product of self-modelling. The experience of free will is caused by the self-model's inability to predict its own actions and, I think, by the consequent endorsement of a given action (the self-model's appropriation of an action as its own). Free will, then, is a label denoting the self-model's interpretation of certain events. Free will, however, does not have an external referent. It originates in the self-model, but does not denote anything beyond the processes constituting the self-model. In a sense, it is analogous to an erroneous pain because it is not caused by actual free decisions of the system, but by the system's interpretations of its own

decision-making processes. The transparency of the underlying mechanism lends itself to the system's misinterpretation of certain events as free choices stemming from deliberative processes.

The experience of freedom is real (just like red qualia and pain qualia are real). It is a real feeling and a real belief and can only be true or false in the same sense that the belief that Sherlock Holmes lives on Baker Street can be true or false, but the belief is a belief about a fictional entity that is not being tracked in the real world. When the computational system attributes free will to itself, the freedom to act, I wish to argue, is not appropriately connected to the environment because it does not denote anything in the environment since freedom of the will is not something that actually exists (it is, as McDermott admits on several occasions, an illusion),[36] it is only the self-model modelling itself as free from causality and this "belief" in the exemption from causality is a false belief, though perhaps unavoidable and indispensable; it is a useful fiction, but a fiction nonetheless.

If free will is just the feeling of freedom, but the mechanical, computational mind never really makes free decisions, then not only are we leading illusory lives, but we construct entire social systems on such fictions. How can anyone be held accountable for anything if everything anyone does is a result of a mechanistic (though very complex and virtually unpredictable, and if Popper is correct, maybe even truly unpredictable) system?

In one sense, it is easy to hold any system (behaviourally) accountable for some X it does. That is, if I see Bob picking flowers in Mrs. Smith's treasured rose

[36] Freedom of the will is an illusion while the transparent deliberative processes responsible for the system's decisions are just as mechanical as those of Rosie the robot.

garden, I can hold Bob accountable in the sense that it was Bob and no other who performed the motions required to pick Mrs. Smith's roses and consequently cause her much grief. It was Bob's visual system that spotted the roses, his hands that picked them, etc. However, there is a problem if I am to pass a moral judgment regarding Bob and his actions.

If, for instance, Bob happens to be a three-year-old boy who does not know any better (who does not understand that picking Mrs. Smith's roses constitutes stealing from Mrs. Smith and moreover, is not acquainted with the notion of stealing and what exactly that entails and why it is a practice that is generally frowned upon), then perhaps condemning Bob for his immoral act is itself condemnable. But, if Bob is a thirty-year-old man who understands the concept of stealing and is aware of the fact that his act may be the cause of Mrs. Smith's heart-attack, then it seems appropriate to resent Bob's actions and in fact blame him for being an immoral man.

What would our intuitions be, however, if Bob (who in all visible ways is a thirty-year-old man) turns out to be a robot running a complex web of programs? And as we head in Bob's general direction in order to chastise him, Bob's designer approaches us and calmly explains the fact that Bob is a robot and goes on to give details about Bob's program, stating that under certain circumstances (which just obtained) Bob will "feel" an irresistible urge to pick flowers. Although the example is simple and quite limited, it should get the main point across. If Bob does not act out of his own free will because Bob is not endowed with a divine-like faculty of will, which allows him to make free choices, then it does not appear to make sense to hold him morally responsible.

If McDermott's theory of mind and the vision of free will it implies is correct, then we are all like Bob. We all lack freedom. However, we also believe that we are free and so, continue our practices of blaming, praising, and punishing in total ignorance of and absolutely oblivious to the fact that our practices of praising and blaming are fundamentally unwarranted[37] and that the objects of our practice of punishment are ultimately victims of circumstance, genetics, etc. If McDermott's view of freedom is right, then what can someone adhering to his theory of mind (along with its metaphysical implications) say regarding moral accountability?

2.5
Non-Moral Judgment and Natural Human Reactions

One possible approach to the above problem would be to consider a consequentialist (and more precisely, a utilitarian) point of view. Our practices of praising, blaming, and punishing, on the utilitarian standpoint, can be viewed as maximizing expected global utility even in the absence of free will because the measure of an action's moral worth is determined by its contribution to overall utility. Thus, whether an agent *S* is free to *A* is not of consequence. Rather, the consequence of *S*'s *A*-ing is that which is of importance. If the action contributes to the overall utility, then it is a morally justified action.

---

[37] Such practices are unwarranted because our commonsense notion of responsibility has it that if an agent *S* is coerced into a certain act *A*, *S* is not to be blamed for *A* because *S* did not intend or will to *A*. Of course, there are certain unintentional actions for which we hold people responsible at least to some degree. For instance, if a drunk driver injures a pedestrian unintentionally, we are inclined to hold the driver responsible for the occurrence because, we reason, it was her intention to drive after having consumed alcohol and that intention is the cause of the unintentional infliction of injury. However, if determinism is the case, then any event, intended or not, will depend on a previous event, which is totally out of the agent's control, having obtained. Thus, the argument goes, the agent is coerced into every single action (that is, both she and the injured party are victims of circumstance and the laws of nature).

On such a view, if we praise, blame, or punish an individual for an act that is out of her control (because she did not act freely), we are justified in our practices (of praising, blaming, and punishing) as long as such practices contribute to overall utility. In fact, if the hedonic calculus works out in such a way that we can get more utility out of the punishment of a clearly innocent individual than out of not punishing her, then we are not only justified, but perhaps even obligated[38] (other things being equal), to go through with the punishment.

Even though a surprising amount of our everyday decision-making may actually run parallel to the utilitarian conception of worth and morality, utilitarianism has some very counterintuitive consequences. McDermott considers the utilitarian option, but quickly rejects it. He writes:

> I find utilitarianism unworkable, for reasons that have been enumerated many times before. Suppose someone proposes to use indigent children as a food item. Utilitarianism suggests adding up the pluses and minuses in order to evaluate the proposal. There's something obviously wrong with a system in which you would even begin this exercise. (McDermott 2001, 228)

Although McDermott does not delve into a lengthy philosophical argument against utilitarianism, his point is well taken. A few other objections raised against utilitarianism in the past (besides the one McDermott alludes to and the problem of replaceability) are: the dilemma of personal integrity, the problem of negative responsibility, and the difficulty with friendship.

I will not concern myself with the many replies these worries have received because the argument against utilitarianism, to be complete (and if one wishes to be

---

[38] Whether we are obligated will depend on many factors, but if the life or well-being of one individual can be exchanged for the lives or well-being of a great number of individuals, *ceteris paribus*, then the utilitarian would be morally condemnable if she did not take the proper course of action (that is, if she did not do everything in her power to ensure the well-being of several people at the cost of sacrificing the well-being of one, all other things being equal).

fair) would go well beyond the present scope of this essay (but I do wish to flag such a possibility even though I do not consider the consequentialist programme to be as compatible with McDermott's view as the Strawsonian approach, which I shall explore in some depth in the next chapter). Since my aim is to enrich McDermott's view and my central concern is the issue of free will, a digression into the vast literature of arguments and counter arguments for and against utilitarianism would take me too far off track. I do wish, however, to briefly ponder J. J. C. Smart's consequentialist conception of responsibility.

Even while assuming that the numerous utilitarian replies (to such problems as replaceability, personal integrity, negative responsibility, friendship, etc.) are inadequate or, at least, that they do not resolve the issues presented above to the satisfaction of non-consequentialists such as McDermott, why can we not adopt the suggestion (put forth by J. J. C. Smart) that we can praise and dispraise, but not judge (in the moral sense) agents for their actions?

Smart distinguishes between two uses of the word praise. The first sense of the word is used when passing non-moral judgments. "Praise and dispraise, in this sense, is simply grading a person as good or bad in some way...Praise and dispraise of this sort has an obvious function just as has the praising of apples" (Smart 1961, 69). The second sense of the word is saturated with moral overtones. This second sense involves the notion of responsibility. Smart argues that if we knew that the thesis of determinism were true, we could compare, grade, praise, and dispraise, but only in the first sense of the word; we could not engage in the practice of praising and blaming with judgements of responsibility attached to such value statements. Smart

writes: "a man's drive [along with his entire nature] is determined by his genes and his environment" (Smart 1961, 71). He concludes by claiming that "[t]he upshot of the discussion is that we should be quite as ready to *grade* a person for his moral qualities as for his nonmoral qualities, but we should stop *judging* him" (Smart 1961, 71).

Smart gives the example of Tommy, who failed to do his homework. I shall alter Smart's example slightly by introducing another character, Jimmy, Tommy's friend, who also failed to do his homework. Tommy, though quite bright, is very lazy while Jimmy, though hard-working, is quite dull. Though it seems intuitive to blame Tommy for not completing his homework (because he opted to play video games rather than do his homework), we do not feel the same reaction to Jimmy's failure to complete his homework (Jimmy sweated over his book for several hours, but could not comprehend the questions being asked). However, if determinism is true, then Tommy's laziness is as much out of his control as Jimmy's dullness. Put somewhat differently, Tommy is causally determined to be lazy (his genes, his upbringing, environment, and other relevant circumstances determine Tommy to be lazy).

Some compatibilists might argue that whereas punishing Tommy for his laziness may actually condition him to be more productive, beating Jimmy will not condition him to be smarter. Thus, punishment can be seen as a necessary formative tool (the same must go for praise and other forms of positive reinforcement), even if it cannot be applied equally in all cases.

Let us imagine that, sometime in the near future, Tommy becomes a law-abiding, productive citizen while Jimmy goes to jail for armed robbery. Jimmy, being

a menace to society, is removed from the social domain, but Jimmy cannot be blamed for what he did or did not do. Locking Jimmy up serves a purpose (i.e. it protects store owners), but it is ultimately unjust (in a deeper moral sense). We can adopt a consequentialist approach to holding individuals responsible even though we refrain from praising and blaming them for their actions.

Knowing that Jimmy is causally determined to be a felon, we do not pass moral judgment on his actions. Jimmy robbed the store, but he could not have done otherwise given his genes, his environment, his history, and the particular circumstances (emotional, physical, environmental, etc.) that presented themselves at the moment he made the decision to walk into a 7-eleven with a gun. The conditionalist[39] might argue that had Jimmy received counselling, he might have made a different choice; or had Jimmy's childhood been different, he would not have had criminal urges, etc. But, I think that the conditionalist account fails.[40] If Jimmy's childhood had been different, for instance, then Jimmy would not be the Jimmy we

---

[39] Some compatibilists maintain that the ability to do otherwise can still be accounted for in a deterministic world. "According to the advocates of this argument—let us call them 'conditionalists'—what statements of the form…S could have done X *mean is*:…If S had chosen to do X, S would have done X" (van Inwagen 1975, 27).

[40] I think that the problem with conditionalist notions of 'could have done otherwise' is the following: if the thesis of determinism were to be true, then, following van Inwagen's definition of 'determinism' (according to van Inwagen (1975), the truth of determinism is contingent on the laws of physics: they must be precise and not probability-driven or statistical), if the conjunction of a certain state of the world A with the laws of physics L entails the state of the world B, then given A and L, B necessarily follows. If the above is true, then it cannot be the case that one possible world P_1 (where the laws of physics consist of the set of laws L), is in state A at time t_1, and state B at t_2, while another possible world P_2 (which is identical to P_1 and where the laws of physics also consist of the set of laws L), is in state A at time t_1, and state D at t_2. Therefore, saying that 'if S had chosen to do X, S would have done X' amounts to saying that although A occurs at t_1 and B comes about at t_2 in P_1, if C had occurred at t_1, then D would have come about at t_2 in P_2. It is like saying that if we run a system governed by a set of laws L, starting it in an initial state A at t_1, it will go into state B at t_2, but if we reset the system and then run it again, starting it in an initial state C at t_1, it will go into state D at t_2. Although the above is true, I do not see how it captures the sense of 'could have done otherwise,' which is necessary for genuine alternative possibilities to be open to an agent.

know and refer to when we speak of him, but someone else (though very similar to Jimmy). Thus, I suspect that Smart would agree that although we may dispraise Jimmy for being a criminal and consequently even lock him up to protect other people from him, we cannot blame Jimmy nor can we hold him morally responsible for being what he is.

From our distant, objective, non-involved standpoint, we may even be inclined to agree with Smart. However, the practice of praising and blaming is not that easy to give up once we find ourselves at the scene of the crime, amid all the turmoil, immersed in the scents and sights of the surroundings with all our senses peaked, our minds alert, gathering data, weeding out the relevant bits of information from a flood of all kinds of background noise, and computing these salient bits of data as quickly as humanly possible. In such cases, there is not much computational power left for philosophical contemplations about free will, responsibility, etc. The store clerk's natural reactions toward Jimmy's deed may include resentment, indignation, dislike, maybe even hatred. And most likely, once Jimmy is arrested and sentenced to a prison term, the clerk will experience some degree of satisfaction (if only because the act of putting Jimmy away produces a restored sense of security in the clerk). But the clerk, even after the fact (while Jimmy is imprisoned) will likely experience a wide array of reactions related to feelings of retribution, blame, etc.

Although the kind of pragmatic approach to praising and blaming suggested by Smart may sound nobler or perhaps more rational than the flood of reactions the clerk might experience, the reactive attitudes of the clerk are natural for, even hardwired into, a great majority of human beings (though there are documented cases

of individuals lacking the natural human emotive capacity). Also, it would be too rash to label such reactive or emotive behaviour as irrational or less rational without having taken a closer look at the nature of such attitudes. In fact, not everyone is taken by such noble pragmatism. Susan Wolf responds quite passionately to the kind of proposal advocated by Smart. She writes:

> The most gruesome difference between this world [the kind of world envisioned by Smart] and ours would be reflected in our closest human relationships—in the relations between siblings, parents and children...spouses and companions. We would still be able to form some sorts of association that could be described as relationships of friendship and love...[But] [w]e would choose friends as we now choose clothing or home furnishings or hobbies, according to whether they offer...the proper combination of pleasure and practicality. (Wolf 1981, 106)

We are finally ready to examine Peter Strawson's notion of reactive attitudes and how the Strawsonian account can enrich McDermott's vision of free will while also delivering a practical, compatibilist version of moral responsibility.

Chapter Three
A Naturalistic Approach to Responsibility

3.1
Reactive Attitudes

Strawson begins with a distinction between "optimists" and "pessimists" about the problem of free will and responsibility. Optimists, in the most basic terms, can be equated with compatibilists who believe that moral responsibility is a coherent notion even if determinism is true. They argue that the practices of praising and blaming do not lose their *raison d'être* if determinism is true. In fact, some optimists claim that the justification of these concepts requires the truth of the thesis of determinism. Pessimists can be loosely equated with incompatibilists (either Libertarians or Hard Determinists). The pessimists argue that if the thesis of determinism is true, then the concepts of moral obligation and responsibility have no application and the practices of punishing and blaming are unjustified. Smart, though not easy to label, can be characterized as a pessimist. Thomas Scanlon also chooses to describe Smart's view as pessimistic (in Strawson's use of the term). Scanlon states that "Smart's analysis is not compatibilist. His aim is to replace ordinary moral judgment, not to analyze it" (Scanlon 1988, 365).

Strawson argues that both the optimists and pessimists misconstrue the facts by over-intellectualizing the issues at stake. However, Strawson's paper is an attempt at reconciling these opposing views. He believes that the optimists are essentially correct, but only if something else is added to their view, namely the notion of reactive attitudes. He also claims that we must "demand of the pessimist a surrender of his metaphysics" (Strawson 1963, 91). I understand Strawson to mean that a

version of compatibilism can account for moral responsibility (although only in light of the notion of reactive attitudes) and that the libertarian metaphysical commitments to alternative possibilities and agent-causal powers or other attempts at providing the agent with self-determination must be abandoned.

Punishment, moral approval/condemnation, etc. are practices or attitudes, which characterize the point of disagreement between the optimists and the pessimists. Strawson proposes to begin by discussing slightly different kinds of attitudes toward others (and of others toward us). Whereas the first kind are attitudes that permit (and even sometimes imply) a certain detachment from the actions or agents which are their objects, the second kind of attitudes are "non-detached attitudes and reactions of people directly involved in transactions with each other...attitudes and reactions of offended parties and beneficiaries...[these are] such things as gratitude, resentment, forgiveness, love, and hurt feelings" (Strawson, P. F. 1963, 75).

It matters to us whether someone's attitude toward us is one of affection, contempt, malevolence, etc. Inter-personal relationships (which may range from the most intimate to the most casual) give rise to reactive attitudes—the non-detached attitudes and reactions towards others and others' actions that are directed at us.

Strawson considers instances where reactive attitudes are natural (where the offended person might naturally or normally be expected to feel resentment). But, he also reflects on cases where special considerations might be expected to modify or mollify this feeling or remove it altogether: (1) unintentional actions are an example of one such case (or instances where an agent could not have done otherwise due to

direct coercion). But none of these types of cases invite us to suspend our ordinary reactive attitudes (we simply excuse the agent even though we may, at first, experience reactions toward them—we excuse them based on the circumstances that reveal the unintentional nature of the act). "They do not invite us to view the *agent* as one in respect of whom these attitudes are in any way inappropriate" (Strawson, P. F. 1963, 77). (2) The second set of excusing conditions is divided into two subgroups: (i) when people are temporally acting out of character (i.e. they are under a hypnotic suggestion, temporally insane, etc.) and (ii) when they are permanently not subject to normal inter-personal relations. Such cases "do not invite us to see the agent's action in a way consistent with the full retention of ordinary inter-personal attitudes and merely inconsistent with one particular attitude. They invite us to view the agent himself in a different light from the light in which we should normally view one who has acted as he has acted" (Strawson, P. F. 1963, 78). Instances of such exemptions include: children, hopeless schizophrenics, etc.

> The second and more important subgroup of cases allows that the circumstances were normal, but presents the agent as psychologically abnormal—or as morally undeveloped. The agent was himself; but he is warped or deranged, neurotic or just a child. When we see someone in such a light as this, all our reactive attitudes tend to be profoundly modified. (Strawson, P. F. 1963, 79)

Strawson distinguishes between *participant attitudes* and *objective attitudes*. Regarding participant attitudes, he states that the natural human commitment to inter-personal human relationships (since we are social creatures) requires inter-personal attitudes. These in turn, require reactive attitudes (except in some cases where we adopt an objective attitude toward individuals who are not members of our moral

community[41] and thus are incapable of the full range of interpersonal relationships we value). Thus, Strawson writes: "This commitment is part of the general framework of human life, not something that can come up for review…[hence] the truth or falsity of a general thesis of determinism would not bear on the rationality of *this* choice" (Strawson 1963, 83). Such reactive attitudes find their analogues in our morality as moral reactive attitudes, which are of a vicarious nature. Such vicarious reactive attitudes are sympathetic, impersonal, disinterested or generalized analogues of the reactive attitudes and deal not so much with resentment as with moral indignation or disapprobation. "They are reactions to the qualities of others' wills, not towards ourselves, but towards others. Because of this impersonal or vicarious character, we give them different names" (Strawson, P. F. 1963, 83), we call them moral reactive attitudes. Strawson emphasizes: "It is not that these attitudes are essentially vicarious—one can feel indignation on one's own account—but that they are essentially capable of being vicarious" (Strawson, P. F. 1963, 84). The moral reactive attitudes apply not only to our demands on others, but also our demands on others for others and on ourselves for others.

We are also capable, according to Strawson, of adopting an objective attitude toward others. Strawson explains:

> The objective attitude may be emotionally toned in many ways, but not in all ways: it may include repulsion or fear, it may include pity or even love, though not all kinds of love. But it cannot include the range of reactive feelings and attitudes which belong to involvement or participation with others in inter-personal human relationships; it cannot include resentment, gratitude, forgiveness, anger, or the sort of love which two adults can sometimes be said to feel reciprocally, for each other. If your attitude towards someone is wholly objective, then though you may fight him, you cannot quarrel with him, and though you may talk to him, even negotiate with him, you cannot reason with him. You can at most pretend to quarrel, or to reason, with him. (Strawson, P. F. 1963, 79)

[41] The "moral community" is what Smilansky (as quoted in a later section) refers to as the "Community of Responsibility."

Strawson writes: "We can sometimes...look on the normal (those we rate as 'normal') in the objective way in which we have learned to look on certain classified cases of abnormality. And our question reduces to this: could, or should, the acceptance of the determinist thesis lead us always to look on everyone exclusively in this way" (Strawson, P. F. 1963, 81)?

Strawson's response to this worry runs along the following lines: being human, that is, being hardwired[42] in such a way that requires inter-personal relationships, we cannot possibly suspend reactive or moral reactive attitudes because they are as deeply ingrained in our nature as human beings as is our need and desire for, as well as inclination and attraction toward inter-personal relationships.[43] Strawson stresses:

> [T]o the...question whether it would not be *rational*, given a general theoretical conviction of the truth of determinism, so to change our world that in it all these attitudes were wholly suspended, I must answer, as before, that one who presses this question has wholly failed to grasp the import of the preceding answer, the nature of the human commitment that is here involved: it is *useless* to ask whether it would not be rational for us to do what is not in our nature to (be able to) do. (Strawson 1963, 87)

Strawson believes that the human commitment to participation in ordinary inter-personal human relationships is too deeply rooted to seriously consider the possibility that a general theoretical conviction (such as the belief in the truth of the deterministic thesis, for instance) would be capable of changing our world in such a way as to abolish inter-personal relationships.

---

[42] Strawson does not actually use the term 'hardwired.'

[43] It may be objected that even though we may not be capable of abandoning the belief that the suspension of reactive or moral reactive attitudes is possible because we are convinced that they are an integral part of our social nature, these reactive attitudes are nonetheless not essential to our human nature. However, insofar as we conceive of ourselves as social beings and insofar as we participate in social frameworks and continue to adhere to social norms, it is unlikely that a massive, long-term suspension of the reactive attitudes is possible or even beneficial. This is, I admit, not a satisfactory reply, but I shall put off the fuller version of this response until a later section where I will attempt to deal with some of the possible objections to Strawson's notion of reactive attitudes.

Even though we are capable of adopting an objective standpoint toward those we rate as normal, such objectivity of attitude is not adopted in response to a general theoretical conviction, but only in response to a particular circumstance. The objective attitude cannot be held indefinitely and in every case. Our practices, Strawson reminds us, do not merely exploit our natures, but they, in fact, express them.

It is quite natural for us to hold on to our reactive attitudes. In *Skepticism and Naturalism: Some Varieties*, Strawson claims that we can no more be reasoned out of our proneness to personal and moral reactive attitudes than we can be reasoned out of our belief in the existence of a body. "[O]ur *general* proneness to these attitudes and reactions is inextricably bound up with that involvement in personal and social interrelationships which begins with our lives, which develops and complicates itself in a great variety of ways throughout our lives and which is, one might say, a condition of our humanity" (Strawson 1983, 33).[44]

Strawson's view introduces two separate standpoints (the participant standpoint and the objective view). Strawson contemplates whether one of these views is the correct standpoint. If the participant point of view is adopted as the true and correct standpoint, then human actions are really proper objects of gratitude,

[44] It may be objected that because natural reactive attitudes towards various groups transform over time (i.e. at least for most educated people today, the reactive attitudes once held toward non-whites—and especially Americans of African descent—or the attitudes toward certain women during the "witch-craze"), such reactions are not stable. It is true that most people (at least in many parts of the "developed" world) no longer believe in witches and thus are no longer disposed to the kinds of reactions their ancestors were so infamously disposed to. However, even though the objects of these reactive attitudes no longer exist, the attitudes are still possible and manifest themselves in various other forms. They no longer target witches, but other objects make themselves available instead. Prejudice, resentment and their like (as well as their opposites) are still very real and active components of our emotive vocabulary. By no means have they been abandoned or diminished, these reactive attitudes are simply put to different uses.

resentment, praise, blame, admiration, hate, etc. If, on the other hand, the objective view proves to be the correct standpoint, then our natural human reactions are nothing more than the natural way in which human beings behave. There is, on this view, no moral reality for these reactions to present or misrepresent and thus, their truth or falsity cannot be questioned. All that exists, on the objective view, is the realm of human behaviour and human reactions to human behaviour. Both are proper objects of study and understanding, but nothing more.

So which is the correct view? Strawson thinks both are the right viewpoints from which we can examine human nature. Strawson writes:

> Relative to the standpoint which we normally occupy as social beings, prone to moral and personal reactive attitudes, human actions, or some of them, are morally toned and propertied in the diverse ways signified in our rich vocabulary of moral appraisal. Relative to the detached naturalistic standpoint which we can sometimes occupy, they have no properties but those which can be described in the vocabularies of naturalistic analysis and explanation (including, of course, psychological analysis and explanation). (Strawson 1983, 38)

There are two faces to Strawson's naturalism. On the soft naturalist account, humans are naturally social beings. Our instinctive commitment to personal and moral attitudes is intimately connected to our inherent commitment to social existence, which arises out of the fact that we are a social species. Hard naturalism suggests that we are capable of sometimes viewing human behaviour in a different light, in an objective or detached manner. The hard naturalist account provides us with the objective standpoint "which involves the partial or complete bracketing out or suspension of reactive feelings or moral attitudes or judgments. To see human beings and human actions in this light is to see them simply as objects and events in nature…in which moral evaluation has no place" (Strawson 1983, 40). We, being rational and intelligent creatures, have the ability to occupy both standpoints. Strawson emphasizes:

> I have suggested that a reconciliation of apparently conflicting views could be achieved by relativizing the conception of the real, of what really exists or is really the case, to different standpoints, acknowledging that a man can occupy one standpoint without rationally debarring himself from occupying the other. (Strawson 1983, 93-94)

Strawson's view strikes me in its similarity to Nagel's distinction between the subjective and objective standpoints. I shall return to this remark when I consider possible objections and replies to the Strawsonian account. But first, there are several other noteworthy points to be contemplated.

3.2
Illusion

Most choices are accompanied by the phenomenal experience of freedom; most of the choices I make (from the trivial to the life-shaping ones) come with the sensation that *I could have done otherwise* and with the feeling that *it was my will that caused my action.* McDermott explains that free will is the self-model modelling itself as free from causality. It may prove useful to explore the illusion of freedom in a little more detail. In his book, *The Illusion of Conscious Will*, the Harvard psychologist Daniel Wegner writes:

> The notion that will is a force residing in a person results in a...problem. Hume...pointed out that causality is not a property inhering in objects...you can't *see* causation in something but must only infer it from the constant relation between cause and effect. Every time the ball rolls into the pins, they bounce away. Ergo, the ball caused the pins to move. But there is no property of causality...hanging somewhere in space between the ball and pins...Causation is an event, not a thing or a characteristic or attribute of an object. In the same sense, causation can't be a property of a person's conscious intention. You can't *see* your conscious intention causing an action but can only infer this from the constant relation between intention and action. (Wegner 2002, 13)

According to Wegner, conscious will can be understood as (1) the experience of performing an action (actions either feel willed or not) or (2) as the causal link between mind and action. One might mistakenly assume that (1) and (2) are the same thing. This mistake, Wegner explains, is the source of the illusion of conscious will.

Wegner puts forth what he calls The Theory of Apparent Mental Causation. "The theory of apparent mental causation, then, is this: *People experience conscious will when they interpret their own thought as the cause of their action*" (Wegner 2002, 64). There are three key sources of the experience of conscious will, according to Wegner. These are: *priority*, *consistency*, and *exclusivity*. "For the perception of apparent mental causation, the thought should occur before the action, be consistent with the action, and not be accompanied by other potential causes" (Wegner 2002, 69).

The Priority Principle states that causal events precede their effects. If X is to be experienced as causing Y, then X must precede Y. Moreover, X cannot occur very long before Y and it particularly cannot take place after Y. For example, if a billiard ball B_1 hits another billiard ball B_2, causing B_2 to roll away, then B_2 must move immediately after (not just some time after and definitely not before) B_1 hits it if B_1 is to be viewed as the cause of B_2's motion. Wegner explains the Consistency Principle by means of the following example:

> When one billiard ball strikes another, the struck ball moves in the same general direction that the striking ball was moving. We do not perceive causality very readily if the second ball squirts off like squeezed soap in a direction that, by the laws of physics, is inconsistent with the movement of the first ball. (Wegner 2002, 78)

And the Exclusivity Principle suggests that people are particularly sensitive to the possibility that there can be other causes besides their own thoughts. "When their own thoughts do not appear to be the exclusive cause of their action, they experience less conscious will. And when other plausible causes are less salient, in turn, they experience more conscious will" (Wegner 2002, 90).

Wegner's view is, I think, compatible with the self-model theory. Wegner states that the experience of will is the way our minds portray their operations to us, but it is not the experience of the actual operation of our minds. In other words, what we experience as free will is what the self-model tells us about our own actions. We feel as though we intend to perform a certain action and thus that the thoughts prior to the action (if they are about that action) are the causes of our action. "We come to think of these prior thoughts as intentions, and we develop the sense that the intentions have causal force even though they are actually just previews of what we may do" (Wegner 2002, 96). Wegner identifies the real causes of our actions with the complex mechanisms that are hidden from consciousness. "We must remember that this analysis suggests that the real causal mechanisms underlying behavior are never present in consciousness. Rather, the engines of causation operate without revealing themselves to us and so may be unconscious mechanisms of mind" (Wegner 2002, 97).

Wegner suggests that actions can "sneak by" without sufficient intentions, but that we correct for such unpleasantly inconsistent actions and confabulate the necessary intentions.

> When life creates all the inevitable situations in which we find ourselves acting without appropriate prior conscious thoughts, we must protect that illusion of conscious will by trying to make sense of our action. We invent relevant thoughts according to the template that conscious agency suggests. (Wegner 2002, 157)

In other words, people justify the things they do. "The process of self-perception is by no means a perfect one; the intentions we confabulate can depart radically from any truth about the mechanisms that caused our behavior" (Wegner 2002, 181).

Wegner explains that the theory of cognitive dissonance holds that people often revise their attitudes in order to justify their actions.

> In a nutshell, the theory says this happens because people are motivated to avoid having their thoughts in a dissonant relationship, and they feel uncomfortable when dissonance occurs. The strongest dissonance arises when a person does something that is inconsistent with a preexisting attitude or desire. (Wegner 2002, 172)

Wegner states that another way of explaining the confabulation of intention is to say we have no attitudes at all prior to an action and that what often happens is that we impute our attitude and the associated intentions after we have acted.[45] Wegner gives an example of such intention confabulation. In a split-brain patient (where information received by the right brain hemisphere was not shared with the left brain and vice versa), "the instruction 'walk' presented to the right brain resulted in the patient's getting up to leave the testing van. On being asked where he was going, the patient's left brain quickly improvised, 'I'm going into the house to get a Coke'" (Wegner 2002, 182). It would appear, Wegner notes, that the instruction "walk" was the cause of the action and the intention of getting a Coke was invented after the fact.

Conscious will, according to Wegner, can only be experienced in the presence of a virtual agent. Without a virtual agent, we would simply not be conscious at all. This view strikes me as being very similar to the self-model approach to consciousness. Wegner explains:

> [V]irtual agents can vary within each person, and perhaps more broadly,...*there is generally a virtual agent for each person*...The development of an agent self in human beings is a process that overlays the experience of being human on an undercarriage of brain and nerve connections. We achieve the fact of having a perspective and being a conscious agent by appreciating the general idea of agents overall and then by constructing a virtual agent in which we can reside [that is, by constructing 'the self']. (Wegner 2002, 269)

---

[45] Kane's notion of *endorsement* (or more precisely, my understanding of Kane's notion of *endorsement*) appears to be a kind of confabulatory process similar to Wegner's idea of confabulation of intention. I will consider Kane's views in the following chapter.

Wegner continues: "*When people project action to imaginary agents, they create virtual agents, apparent sources of their own action. This process underlies spirit possession and dissociative identity disorder as well as the formation of the agent self*" (Wegner 2002, 221).[46]

Wegner draws the analogy between conscious will and a compass used for steering a ship. He claims that just as the compass readings do not steer the ship, the conscious experience of will does not cause human actions. But, as Wegner makes clear, the conscious will is an indicator to which we refer as we steer. He writes: "the occurrence of conscious will brands the act deeply, associating the act with self through feeling, and so renders the act one's own in a personal and memorable way. Will is a kind of authorship emotion" (Wegner 2002, 325).

> Conscious will is particularly useful, then, as a guide to ourselves. It tells us what events around us seem to be attributable to our authorship. This allows us to develop a sense of who we are and are not. It also allows us to set aside our achievements from the things that we cannot do. And perhaps most important for the sake of the operation of society, the sense of conscious will also allows us to maintain the sense of responsibility for our actions that serves as a basis for morality. (Wegner 2002, 328)

As Smilansky observes, "if libertarian assumptions *carry on their back* the CC [Core Conception][47] distinctions, which would not be adhered to sufficiently without them, an illusion which defends these libertarian assumptions seems to be just what

---

[46] Wegner provides the following anecdote as an example: "One day, a visitor came into Bergen's [a ventriloquist's] room and found him talking—not rehearsing—with Charlie [his dummy]. Bergen was asking Charlie a number of philosophical questions about the nature of life, virtue, and love. Charlie was responding with brilliant Socratic answers. When Bergen noticed that he had a visitor, he turned red and said he was talking with Charlie, the wisest person he knew. The visitor pointed out that it was Bergen's own mind and voice coming through the wooden dummy. Bergen replied, 'Well, I guess ultimately it is, but I ask Charlie these questions and he answers, and I haven't the faintest idea of what he's going to say and I'm astounded by his brilliance (Siegel 1992, 163)'" (221). Other examples of virtual agency are: imaginary agents (imaginary friends), spirits, mediums, possessions, multiple personalities, etc.

[47] The Core Conception is the elementary ethical conception that takes as its focus the necessity of considering free will as a prerequisite for morality.

we need" (Smilansky 2000, 173). That is, the belief in free will may, in fact, be responsible for our acting morally much of the time.

There are other advantages of being in possession of the illusion of conscious will. Wegner summarizes the results found by Langer and Rodin, in their 1976 and 1977 studies, according to which elderly people given new control opportunities (even as insignificant as being responsible for watering a plant) show renewed resilience in psychological and physical well-being. While Bulman and Wortman, in a 1977 study, report that victims of paralysing accidents who believed that their victimization was their own responsibility were better equipped to cope with their misfortune. Wegner comments: "the habit of taking responsibility seemed to carry over from the accident into the pursuit of adjustment in the aftermath...it is reasonable for a person who perceives control in one area to suspect the possibility of such control in another" (Wegner 2002, 330).

The illusions of apparent mental causation, according to Wegner, are the building blocks of human psychology and social life. "It is only with the feeling of conscious will that we can begin to solve the problems of knowing who we are as individuals, of discerning what we can and cannot do, and of judging ourselves morally right or wrong for what we have done" (Wegner 2002, 342). The illusion of freedom, then, endows us with dignity and it is that which makes us human.

3.3
Reactive Attitudes and Illusionism

In "Free Will and Respect for Persons," Smilansky argues that treating someone as a responsible person does not depend on the truth or falsity of

determinism. What is of importance, however, is the acquisition of certain capacities (capacities for awareness, deliberation, choice, and intentional action), which enable an adult person to act responsibly. On the Strawsonian view, a baby does not belong to the "Community of Responsibility" (neither do kleptomaniacs, etc.), but the "normal" adult person does. It is irrelevant whether the normal adult human being is determined or not. The salient point is that she is capable of being held responsible by her peers in the "Community of Responsibility." Smilansky reminds us, however, of the ultimate arbitrariness of all moral judgments:

> While membership in a Community of Responsibility permits punishment of the guilty student, it at the same time forbids 'punishment' of the innocent one. Nevertheless, the actions of the drug dealer [the guilty student] were, in one way, merely an unfolding of the given, of matters that, causally constituting her, were ultimately beyond her control. Together with the moral obligation to respect and to track (in our own reactions and practices) identity, choice, and responsibility, we must also not forget the ultimate *arbitrariness* of it all. (Smilansky 2005, 256)

How, Smilansky asks, can we begin to reconcile the arbitrariness found at the fundamental level with the natural reactions we experience toward others as well as toward ourselves? The adoption of the objective standpoint, which exposes the lack of control over our own actions, threatens to affect our reactions to others and ourselves and the evaluations of performance we practice on a daily basis.

Smilansky argues that Strawsonians should adopt illusionism as a tool for coping with the above problem. An emphasis on the respect for persons, Smilansky admits, is difficult to reconcile with the perhaps somewhat demeaning talk of the positive benefits of illusion. But, Smilnasky continues, such a state of affairs where we would not need recourse to illusion, but determinism would still be true carries a price we cannot afford. Such complete knowledge of the truth of the lack of control over our own choices would, according to Smilansky, "put our moral house at grave

risk" (Smilansky 2005, 257). He continues: "The moral house we have is essentially a Community of Responsibility…In short, the ethical importance of the Community of Responsibility should be taken very seriously, but the ultimate perspective threatens to *present* it as a farce, a mere game without foundation" (Smilansky 2005, 257). Thus, the fact that we are endowed with the illusion of freedom is actually a blessing rather than a demeaning feature of our limited humanity. In fact, Smilansky argues that illusionism is necessary for the Strawsonian notion of reactive attitudes.

> Respect for persons requires on the one hand respect for agency, the establishment of a moral order based on responsibility, and the attempt at human empowerment within compatibilist spheres; on the other hand, it requires recognition of the limitations and shallowness of these spheres, where everything that goes on is ultimately an unfolding of the given, beyond anyone's control. This dissonance already calls for illusion to serve a 'functional' role, that of safeguarding the partly valid compatibilist-level 'form of life' (a primary condition for respect for persons) from the threat of the ultimate hard deterministic perspective that levels all of us. But beyond the 'functional' stage lies the 'existential' stage, where philosophically we can recognize how intimately our fundamental evaluations of ourselves and of others, and of our reactions to one another, depend on the false libertarian picture. We confront the deep dangers of awareness and internalization of the truth. At the depths, the libertarian illusion is constitutive of our very humanity; it is a condition for deep self-respect and for respect for persons. (Smilansky 2005, 260-261)

Whether illusion is, in fact, necessary for Strawson's account is debatable, but in light of McDermott's self-model theory of mind, it is at least safe to assume that illusionism is compatible with the Strawsonian approach to responsibility.

3.4
Frankfurt's Compatibilism

The commonsense notion of freedom (the feeling of ultimate control over one's actions) appears to remain regardless of the amount of evidence supporting a contrary possibility (that human beings are collections of events and human actions mere occurrences—causes and effects embedded within long, complex causal webs). The intuitive belief in free will originates, I suspect, in light of introspective evidence. That is, when I think of myself as an agent in the world, I cannot but see myself as

free. Furthermore, I recognize the fact that my actions usually (or at least quite often) are the result of my own reasoning, which distinguishes me from the merely "physically" causal nature of rocks and other inanimate objects that surround me at all times. Searle captures this experience of freedom quite eloquently:

> We know we could have done something else, because we choose one thing for certain reasons. But we were aware that there were also reasons for choosing something else, and indeed, we might have acted on those reasons and chosen that something else...it is just a plain empirical fact about our behaviour that it isn't predictable in the way that the behaviour of objects rolling down an inclined plane is predictable...If we want some empirical proof of this fact, we can simply point to the further fact that it is always up to us to falsify any predictions anybody might care to make about our behaviour...that sort of option is simply not open to glaciers moving down mountainsides. (Searle 1984, 87-88)

However, no matter how free we may feel, it is quite possible, maybe even likely, (especially in light of the contemplations of previous sections) that our experience of freedom may be illusory.

Conditionalism, to my mind, encounters the previously mentioned problems due to its stubborn adherence to the libertarian notion of responsibility, which requires alternative possibilities. Harry Frankfurt has, I think, successfully shown that the notion of alternative possibilities does not actually play a role in the formulation of moral judgments and our practices of praising and blaming. I think that a more sophisticated form of compatibilism must, as Frankfurt's does, abandon all libertarian preconceptions and deal with responsibility independently of the question of whether metaphysical freedom actually exists.

Frankfurt claims that the Principle of Alternative Possibilities (PAP) is false; a person can be morally responsible even if she could not have done otherwise. According to Frankfurt, a person can be in a situation where she cannot do otherwise, but where these circumstances do not actually impel (coerce) her to act or in any way produce her action. Moral responsibility, on Frankfurt's account, can only be

suspended if the agent is coerced. The doctrine that coercion excuses the agent from responsibility is not correctly understood when viewed as a version of PAP.

Frankfurt explains:

> [I]t is incorrect to regard a man as being coerced to do something unless he does it *because of* the coercive force exerted against him...When we excuse a person who has been coerced, we do not excuse him because he was unable to do otherwise. Even though a person is subject to a coercive force that precludes his performing any action but one, he may nonetheless bear full moral responsibility for performing that action. (Frankfurt 1969, 171)

Frankfurt illustrates his point by means of a now famous thought experiment. There are many versions of this experiment and I shall paraphrase one such version (one that makes Frankfurt's argument against PAP very clear).

The Frankfurt-style case goes as follows: Jones is sitting on a rooftop as he waits for the right moment to assassinate Smith. Black (an evil neurosurgeon) also desires Smith's death and wants Jones to kill Smith. Jones does not know that Black wants him to kill Smith. Jones has his own reason for assassinating Smith. Black, in order to ensure that Jones goes through with the assassination (just in case Jones has a last-minute change of heart and decides against killing Smith), implants a device into Jones' brain, which, if activated by Black, will manipulate Jones' neural states and will override his decision not to kill Smith in favour of a forced decision to commit the murder. So, in essence, if Black makes use of the device, Black can be seen as applying a direct form of coercion and thus, if Black uses the device to force Jones to murder Smith, Jones will not be responsible for Smith's death. But, Black decides to use his device only if absolutely necessary and so, the device is set to activate (thereby triggering Jones's decision to kill Smith) only if it detects Jones changing his mind. Otherwise, the device will remain inactive. Thus, no matter what happens, Jones will end up killing Smith and so, it is true that Jones could not have done

otherwise. As the story goes, Jones does not decide to abandon his murderous plan and goes ahead with the successful assassination of Smith (without the intervention of Black's device). In such a case, even though Jones was truly unable to do otherwise, we judge Jones to be blameworthy for his act because Jones willed it and his decision to kill as well as his act of murder was not coerced.[48]

One might object that if determinism is true, then one's decision to do X is just as determined and inevitable as one's will to X and one's act of X-ing. "The revised principle of alternate possibilities will entail, on this assumption concerning the meaning of 'could have done otherwise', that a person is not morally responsible for what he has done if it was causally determined that he do it" (Frankfurt 1969, 175). Frankfurt does not think this revised principle is acceptable because even if causally determined, if S were to do X regardless of the fact that S cannot do otherwise, S is morally responsible for doing X since, were it the case that S could do otherwise, S would have done X anyways. Frankfurt continues:

> The following may all be true: there were circumstances that made it impossible for a person to avoid doing something; these circumstances actually played a role in bringing it about that he did it, so that it is correct to say that he did it because he could not have done otherwise; the person really wanted to do what he did; he did it because it was what he really wanted to do, so that it is not correct to say that he did what he did only because he could not have done otherwise. Under these conditions, the person may well be morally responsible for what he has done. (Frankfurt 1969, 176)

Frankfurt-style examples attempt to prove that alternative possibilities are not necessary for responsibility. I think that when we talk about free will (in normal circumstances and not while engaged in philosophical discussion), we have some

---

[48] A more compelling example, which brings out Frankfurt's intuition, is the case of the willing addict who would do whatever it takes to reinstate the grip of his/her addiction were it to somehow weaken. "The willing addict's will is not free, for his desire to take the drug will be effective regardless of whether or not he wants this desire to constitute his will. But when he takes the drug, he takes it freely and of his own free will…His will is outside his control, but, by his second-order desire that his desire for the drug should be effective, has made this will his own" (Frankfurt 1971, 335-336).

version of libertarianism in mind. For the conception of free will we employ in everyday discourse to be true, some genuine alternatives must be open to us. As we have seen, however, determinism does not seem to be a very hospitable ground for the libertarian notion of freedom.

Thus, I read Frankfurt's argument as successfully separating the notion of metaphysical freedom from the problem of responsibility. That is, I understand Frankfurt's argument as stating that even if one does not have free will (that is, even if determinism reigns supreme), one can still be held responsible for one's actions because when we talk about responsibility, we are discussing a matter that is not dependent on the existence of metaphysical freedom.

What, then, makes one's act one's own? The answer to this question will need to say something about the quality of one's will.[49] Frankfurt distinguishes between first-order and second-order desires. First-order desires are desires over which one does not exercise any control. These may include such things as being hungry and thus desiring food, etc. Second-order desires are those desires that have desires as their objects. In other words, desiring to have a certain desire constitutes a

---

[49] 'Quality of Will' is, according to McKenna, a notion that is shared by both Frankfurt and Strawson. McKenna argues for this in: McKenna, M. (2005). "Where Frankfurt and Strawson Meet." *Midwest Studies in Philosophy: Free Will and Moral Responsibility*, Vol. 29. Ed. Peter A. French and Howard K. Wettstein. Guest Ed. John Martin Fischer. Malden, MA: Blackwell Publishing. 163-180. McKenna writes: "Suppose that a person tells us that he did what he did because he was unable to do otherwise…We understand the person who offers the excuse to mean that he did what he did only because he was unable to do otherwise…And we understand him to mean, more particularly, that when he did what he did it was not because that was what he really wanted to do…So as I see it, Frankfurt's treatment of the excuse 'I could not have done otherwise' is the precise point where he meets Strawson. Strawson's account would not have it that this excuse could work in the absence of considerations that would prove that a person did not act from a morally objectionable will…There is yet a deeper point beyond the mere fact that Frankfurt and Strawson treat the pertinent excuse in a similar way. It is this point I have been most concerned to demonstrate. Both Frankfurt and Strawson join company on a fundamental insight about moral responsibility. This is brought forth in Frankfurt's point regarding the irrelevance of Black's presence, and in Strawson's Quality of Will Thesis. What matters is what a person does do and what quality of will motivates her doing it, not what she could have done" (McKenna 2005, 175-176).

second-order desire. For instance, one may experience a first-order desire of wanting food without the second-order desire of wanting to want food (if, for example, one is fasting and does not want to eat even though one is, in fact, very hungry).

What do first and second-order desires have to do with quality of the will? In and of themselves, first and second-order desires do not have very much in common with quality of the will. What Frankfurt regards as essential to being a person are second-order volitions, which differ somewhat from second-order desires. If someone has a second-order volition, she wants a certain desire to be her will. This is what "quality of the will" refers to. If Jones wants to kill Smith and if he has a second-order volition to kill Smith (that is, if Jones wants his desire to kill Smith to be his will), then, if Jones succeeds in killing Smith, regardless of whether he could or could not have done otherwise, he is responsible for the deed.[50] Jones is responsible for the murder in purely behaviourist terms: it is Jones himself (and no other) who performs the deed. But Jones is also responsible in psychological terms: it is Jones' will to kill Smith (in fact, it is Jones' will to will to kill Smith).

Thus, from a third-person point of view, it is quite clear that it is Jones who is responsible for the act of killing Smith (someone can actually observe Jones pulling the trigger and assassinating Smith). Also, Jones too, upon introspection (from the first-person point of view), will see himself as the one responsible for the murder because his own second-order volition causes him to kill Smith.

---

[50] If we know that Jones wants his desire to kill Smith to be his will, we naturally experience emotive reactions (reactive attitudes) toward Jones and feel compelled to blame him for his action. I think that when we evaluate whether someone is responsible for an action, we take the quality of their will into consideration. The quality of will appears to be of significance on Strawson and Frankfurt's views alike.

Moreover, "having the freedom to do what one wants to do is not a sufficient condition of having a free will. It is not a necessary condition either. For to deprive someone of his freedom of action is not necessarily to undermine the freedom of his will" (Frankfurt 1971, 331). Such a person may simply be unable to translate her desires into actions.

There may also be third, fourth, and n^{th} -order volitions. How does one escape such infinite regress? If a person identifies with her first-order desire (via a second-order volition), Frankfurt argues, it becomes unnecessary to have higher-order volitions beyond this second-order level "[w]hen a person identifies himself *decisively* with one of his first-order desires" (Frankfurt 1971, 332).[51] Also, higher-order volitions do not need to be formed deliberately or as a result of a struggle to ensure that they are satisfied. They may be much more thoughtless and spontaneous than this.

Frankfurt declares: "It is not true that a person is morally responsible for what he has done only if his will was free when he did it. He may be morally responsible for having done it even though his will was not free at all" (Frankfurt 1971, 334).

One might object that given the truth of the thesis of determinism, it is never anyone's will to have a given second-order volition. If determinism is the case, then whether I want to will X is itself determined and beyond my control. And so, if I did

[51] Frankfurt explains: "The decisiveness of the commitment he [the agent] has made means that he has decided that no further question about his second-order volition, at any higher order, remains to be asked. It is relatively unimportant whether we explain this by saying that this commitment implicitly generates an endless series of confirming desires of higher orders, or by saying that the commitment is tantamount to a dissolution of the pointedness of all questions concerning higher orders of desire" (Frankfurt 1971, 333).

not choose to have a certain will, how can my action be assigned to me (how can I be held responsible for an action) based on the fact that I willed the action?

This is, I admit, a problem with any view that assumes, as part of its account, the truth of determinism. However, as far as I understand Frankfurt's argument, freedom has nothing to do with responsibility. The question of whether human beings possess metaphysical freedom can be answered in the negative while the question of responsibility can be answered affirmatively. Jones could not have done otherwise. Jones was determined to kill Smith. The difference is that Jones, in the actual world, is determined by his own will while in the nearby possible world where Black activates the device Jones is determined by Black's will. Moreover, if Jones could have done otherwise in the actual world (where he is determined by his own will), he still would have assassinated Smith. In other words, knowing that Jones not only authored, but also reflectively endorsed his action invokes various moral reactive attitudes because Jones' quality of will (or his character) is judged as being vicious or malicious in nature.

Responsibility, then, on Frankfurt's view, is intimately tied to higher-order volitions, but it is not necessary for these higher-order volitions to come about by an agent-causal power. I think that this view, though useful in many ways, must ultimately succumb to Smilansky's criticism of all compatibilist accounts. At the fundamental level, there appears to be arbitrariness in the assignment of responsibility. I suspect that this arbitrariness will be present in all non-libertarian accounts, but I also think that the distinction between the objective and subjective standpoints (a distinction recognized by both Strawson and Nagel) will prove

significant. I shall return to this point shortly, but for the time being, I feel compelled to examine several objections to the Strawsonian picture.

3.5
Reactive Attitudes (Objections and Replies)[52]

One quite important objection to Strawson is that it may be possible that reactive attitudes are simply not central to human nature and that like other attitudes human beings had in the past (for instance, certain superstitious attitudes arising from pre-scientific explanations of physical phenomena), these too can and will dissipate if the thesis of determinism (along with a mechanistic view of mind) proves to be true. The question, then, is whether reactive attitudes are truly inherent in human nature.

Emotions, like our senses, on the evolutionary account, play a very important role in our individual lives as well as in the survival of our species. Such emotions like love and many types of altruistic tendencies perform complex and essential functions for us as individuals and as a species.

Owen Flanagan supports Strawson's claim about reactive attitudes. In *The Problem of the Soul* he states: "certain emotions are universal and what I call 'proto-moral'—meaning that long before *Homo sapiens* articulated such things as moral codes, we used our emotions to regulate social life" (Flanagan 2002, 302). He cites Charles Darwin and Paul Ekman's lists of pre-lingual human emotions: fear, anger, surprise, happiness, sadness, disgust, and contempt as well as sympathy, fidelity, and courage. Reactive attitudes such as resentment, like the other primal emotions, he

[52] Some of the arguments presented in this section were originally developed in my paper, "Hardwired Freedom: In Defence of Strawson's Reactive Attitudes," written for Associate Professor and Chair Bruce Hunter in partial fulfillment of the requirements for his Free Will seminar held in the Fall Term of 2005.

claims, can be linked directly to the list provided by Darwin and Ekman. He continues: "Now the Darwinian will say that any universal trait is probably a biological adaptation. But the critic will rightly point out that this only tells us that when the trait evolved it led to reproductive success…This much tells us why Mother Nature endowed us with the relevant equipment" (Flanagan 2002, 307). And this may be enough to support Strawson's claim that reactive attitudes are a natural aspect of human nature, one that evolved over time because of its usefulness and survival value.

However, the question of whether these attitudes are adaptable still remains. Flanagan sheds light on the question of adaptability by considering the glossy leaves of the eucalyptus tree that have adapted to preserve moisture, which is beneficial to its survival in an arid climate. He states that if the climate were to change to a tropical one, then "the adaptation subserving moisture retention is no longer fitness enhancing, and eucalyptus trees will become oversaturated with retained water and rot away" (Flanagan 2002, 311). He further writes:

> This example's relevance to the case of the proto-moral emotions is straightforward. Even if these emotions were selected for and maintained in the species because they were once adaptive, this does not establish that they remain adaptive in the environmental niches we now occupy. (Flanagan 2002, 311)

Thus, our emotions and reactive attitudes are quite static and we are, as it were, stuck with them.

Though they are not adaptive, the reactive attitudes may still be seen as potentially modifiable. Flanagan states that "Strawson himself concedes that the reactive attitudes are subject to forces of cultural learning" (Flanagan 2002, 312). He continues:

> We cannot disentangle emotions from morality, nor should we want to. But we can moderate and adjust our emotions, making them more "apt" to different situations, different social environments, different moral conceptions. (Flanagan 2002, 313)

Thus, although our emotions and our reactive attitudes are static (in other words, they are, for the most part not disposable), these same reactive attitudes that are such an essential aspect of human nature may be moderated to fit different situations (such as, for instance when we deal with a child as opposed to an adult), different social environments (perhaps the type of resentment one may feel toward the members of the opposing team when playing soccer or football, for instance, though still exhibiting a natural reactive attitude, is adjusted, I would assume, to the circumstance of the social activity/environment),[53] and different moral conceptions (although all human beings share these reactive attitudes, they may be applied to one type of situation in one moral community and a totally non-related circumstance in another moral community). Regardless of cultural differences, however, I am compelled to argue that since reactive attitudes are an essential and inseparable aspect of human nature, the notion of responsibility and thus the passing of moral judgment, which arises out of them is as natural a phenomenon as the reactive attitudes, which we judge to be the point of origin of such concepts as responsibility and practices such as the passing of moral judgment. Therefore, even if the thesis of determinism were true, and even if we were to become aware of the fact that the nature of the world is determined, our reactive attitudes, as well as the notion of responsibility and the practice of judgment that go along with the reactive attitudes, would not be suspended.

---

[53] Although sometimes such "adjusted" reactive attitudes do get out of control. Violent episodes exhibited by some soccer fans are good examples of the fine line between modified ("adjusted") reactions and regular reactive attitudes being crossed. In these cases, the adjusted reactions are elevated to actual feelings of hatred, resentment, etc.

Susan Dwyer, in an attempt to reaffirm Strawson's claim regarding the natural status of reactive attitudes, cites several studies, which appear to confirm the innateness of such attitudes. For instance, she explains that "children, within the first year of life, manifest a range of pro-social behaviors, like helping, comforting, and sharing" (Dwyer 2003, 188). This implies, she argues, that human beings are innately empathetic. She continues by stating that very young children are normatively sensitive, meaning that young children care about standards of various kinds. "In particular, they discern a difference between moral and conventional rules" (Dwyer 2003, 188). Dwyer also points out that research meant to track the emotional attributions children make to agents reveals that children are sensitive to others' distress. In order to strengthen Flanagan's argument, I quote Dwyer once again: "Empirical research in moral development…seems to reveal that moral development is just what the Strawsonian model would predict" (Dwyer 2003, 191).

Also, as Daniel Dennett makes clear, the issue at hand is not whether determinism is true, but whether the knowledge of the truth of determinism would have an impact on our reactive attitudes. I quote Dennett: "After all, if determinism is true now, it always has been true…Modern science isn't *making* determinism true, even if it is discovering this fact, so things aren't going to get worse, unless it is believing in determinism rather than determinism itself that creates the catastrophe" (Dennett 1984, 15). And, as evident from the discussion above, such knowledge should not be able to undermine our essential human instincts.

It might be objected that the reactive attitudes can be suspended much like our "carnivorous nature" can be overridden[54] (as in many cases it is when people choose to become vegetarians). It is quite true, and Strawson would be the first to admit it, that reactive attitudes can be suspended. However, unlike making the choice of replacing one's meat-eating habits with a vegetarian preference, the reactive attitudes are intimately connected to the self-model, which, as already mentioned, cannot come up for review.

The self-model is responsible for the subjective point of view, which causes agents to, among other things, have the belief that one is free from causality and thus, the subjective view gives rise to the illusion (or experience) of freedom. This is not a feature that can be abolished even though such a belief can be suspended for the duration of deep, philosophical reflection upon the matter. The reactive attitudes stem in part from this constraint and thus, when in the grip of the subjective point of view, one is necessarily blind to the objective standpoint. The reactive attitudes, being an inseparable aspect of the subjective viewpoint, can only be questioned in moments of reflection where we allow ourselves to adopt the point of view of the universe, as it were, but, as already argued, such reflections come abruptly to an end as soon as we return to the subjective standpoint, which we must because our self-models generate it. And because we are our self-models (that is, we identify ourselves with the content of our self-models), the reactive attitudes are an integral part of ourselves.

It is true that we suspend our reactive attitudes quite frequently. For instance, for professionals (like psychoanalysts), such suspensions are an important part of

[54] I would like to thank Professor Wesley Cooper for pointing out this analogy.

their vocation. Similarly, a surgeon is expected to adhere to the Hippocratic Oath (which states, among other things, that the physician will never deliberate harm to anyone for anyone else's interests) even in the terrible circumstance where the surgeon does, in fact, wish harm upon her patient (because, for instance, the patient is a notorious murderer). In fact, any physician not adhering to the Oath is morally condemnable.

The question, then, is whether, if we are capable of suspending our reactive attitudes even in extreme cases, why can we not integrate such suspension of reactive attitudes into all aspects of our lives? My guess is that we could not because, if McDermott is correct about the status of our self-models (that they comprise our self-knowledge and that they are responsible for how we view and interact with the outside world) our subjective experience of the world cannot be abolished. It is important to remember that the content of the self-model does not have to be true to be useful. This subjective point of view is composed, in part, of the subjective experiences of various emotions, moods, etc. caused by the physiological processes in our bodies. These, in turn, contribute to the force with which our reactive attitudes grip us in certain circumstances. Perhaps we could give up our reactive attitudes, but this would entail a dramatic restructuring of our self-models and a superhuman control over the emotional nexus that is hardwired into us. And perhaps this is a possibility. Many cultures tell stories of enlightened beings with superhuman patience and a divine sense of right and wrong and we sometimes hear of saintly individuals who harbour no ill feelings toward their prosecutors and executioners. When the world is such that all (or at least the majority of) creatures reach such

heights of detachment and selflessness, then perhaps the Strawsonian analysis will no longer ring true, but as it stands, those we label as members of our moral community all exhibit reactive attitudes and are capable of suspending them only temporarily.

However, according to David Silver,[55] even though moral responsibility arises from our reactive natures, once the concept of moral responsibility is in place, it will remain even if the reactive attitudes are permanently suspended. Thus, even the enlightened Buddhist, after having transcended the illusory self and thus the illusion of freedom along with the accompanying reactive attitudes, can remain a moral being and is capable, if she so wishes, to partake in our moral community (such a being can at least continue to be praised and admired by the non-enlightened members of the moral community).

Another possible problem for Strawson is one Watson brings to the surface and Michael McKenna summarizes. McKenna writes: "If a condition of responsible moral agency is membership within the moral community, then those individuals who are so evil that their behavior eschews the values of moral community altogether should fail to count as moral agents, and hence, fall outside the scope of our ascriptions of responsibility" (McKenna 1998, 127). In other words, it seems as though, on the Strawsonian account, evil excuses itself. That is, since someone who commits evil acts is not part of the moral community in virtue of the evil act committed and if the practice of praising and blaming requires membership in the moral community, then those who commit evil acts are excluded from the moral community and so, exempt from blame and thus cannot be held responsible.

---

[55] Silver, David. (2005). "A Strawsonian Defense of Corporate Moral Responsibility." *American Philosophical Quarterly*, Vol. 42, No. 4. 279-293.

McKenna argues that moral responsibility is not constituted by membership in the moral community alone, but that "Strawson's basic naturalistic insight can be preserved by explaining responsible moral agency in terms of *capacity* for membership within the moral community" (McKenna 1998, 129). He concludes: "An individual is a competent moral agent so long as she has the *capacity* to participate within the moral community; she need not *actually* be a member. If she does have the capacity, she stands within the scope of our moral address" (McKenna 1998, 142). The *capacity* to participate within the moral community can be understood in terms of Paul Russell's notion of capacity as outlined in his paper "Responsibility and the Condition of Moral Sense."[56] He writes that an agent has the capacity to partake in the moral community if she is capable of internalizing the reactive attitudes. That is, if she is capable of experiencing the relevant emotions and if she accepts the legitimacy and significance of the considerations that produce these feelings (as opposed to the agent who has a merely external attitude toward these sentiments). A psychopath, being intelligent, may act quite normally since she understands the consequences (i.e. the resulting practices of praise and blame) of her actions, but lacks the ability to feel or internalize the reactive attitudes.

The issue of how we can tell whether someone's action was indeed intentional or whether someone is in fact a psychopath arises because it is quite conceivable that we may be mistaken in our judgments about the appropriate circumstances for the suspension of our reactive attitudes. Jonathan Adler argues, however, that "[t]he

[56] Russell, P. (2005). "Responsibility and the Condition of Moral Sense." *Philosophical Topics*. Ed. John M. Fischer.

question for the reactor is whether his judgment (or action) is justified, not whether it cannot be wrong" (Adler 1997, 904).

Thomas Nagel also objects to Strawson's account. In "Freedom,"[57] he argues that the feeling of autonomy stems from the fact that we view the world from the inner (subjective) viewpoint. The subjective standpoint results in the common notion of autonomy. Nagel describes our ordinary conception of autonomy as the belief that antecedent circumstances leave certain actions undetermined and that these undetermined actions are determined only by our choices. The problem with such explanations, however, is that "an autonomous intentional explanation cannot explain precisely what it is supposed to explain, namely *why I did what I did rather than the alternative that was causally open to me*" (Nagel 1986, 235). Thus, "[a]t some point this question will either have no answer or it will have an answer that takes us outside of the domain of subjective normative reasons and into the domain of formative causes of my character or personality" (Nagel 1986, 235).

The domain of formative causes of one's character or personality is the domain of what Nagel calls the objective view. This inevitable transition from the subjective to the objective view leads to the erosion of the notion of autonomy and thus to the vanishing of inter-personal attitudes and eventually to the evaporation of the notion of responsibility.

---

[57] Nagel, T. (1986). "Freedom." *The View from Nowhere*. New York, NY: Oxford University Press. 110-137. Also published in Gary Watson's anthology: Nagel, T. (1986). "Freedom." *Oxford Readings in Philosophy: Free Will, 2nd Ed.* Ed. Gary Watson. Oxford: Oxford University Press. pp. 229-256. 2004.

Nagel claims that we adopt the objective view in an attempt to make more informed and freer choices, but the objective view, when taken too far, results in a view of ourselves as aspects or parts of the entire system. He writes:

> Something peculiar happens when we view action from an objective or external standpoint. Some of its most important features seem to vanish under the objective gaze. Actions seem no longer assignable to individual agents as sources, but become instead components of the flux of events in the world of which the agent is a part. (Nagel 1986, 229)

So, the objective view, though intended to make more informed (free) choices, robs us of the "illusion" of autonomy.

Nagel criticizes Strawson's account, which states that reactive attitudes are a given fact of human society and that, as a whole, this framework of attitudes, neither calls for, nor permits, an objective or rational justification. According to Nagel, however, the external point of view does undermine our practices of praising and blaming because the objective standpoint precludes the possibility of projecting into another's point of view and such a practice is requisite for judgments of responsibility.

Be that as it may, Nagel does not hold, however, that we are capable of adopting the objective view indefinitely. "The bafflement of moral judgments by objective detachment is unstable" (Nagel 1986, 242). He continues with the William Calley example:

> We may be able temporarily to view William Calley, for example, as a phenomenon—a repulsive and dangerous bit of the zoosphere—without condemning him on the basis of a projection into his standpoint of our own sense of genuine alternatives in action. But it is next to impossible to remain in the attitude of inability to condemn Lieutenant Calley for the murders at My Lai: our feelings return before the ink of the argument is dry. (Nagel 1986, 242)

It would appear, then, that although the objective point of view is conducive to the suspending of our reactive attitudes, we are incapable of retaining such an external

standpoint for very long; we are bound, inevitably, to slip back into the subjective view, which resurrects our notion of autonomy as well as our ability to project this notion onto others, our conception of responsibility, and the reactive attitudes that accompany the subjective view and give rise to our practices of praising and blaming.

Though I do not have sufficient space to get into the intricacies of Paul Russell's paper *Strawson's Way of Naturalizing Responsibility*, I propose to offer a sketchy paraphrase of his concern in order to bring to light yet another serious objection to Strawson's notion of reactive attitudes.[58]

Russell worries that since we can suspend our reactive attitudes in some cases (i.e. when dealing with children or the insane), we are capable of suspending them in all cases if we deem all cases to be morally incapacitated. And, the truth of the thesis of determinism would prove to be a case where all agents would in fact become morally incapacitated. To rearticulate my sketchy paraphrase in Russell's own words:

> Strawson acknowledges that we may find ourselves in circumstances where our reactive attitudes are not called for or are inappropriate...Clearly, then, while we may remain prone to reactive attitudes, they are, with us, in these circumstances, wholly inactive and disengaged (because they are acknowledged to be inappropriate and uncalled for)...the Pessimist claims only that we can and must cease to entertain reactive attitudes toward any and all individuals who are morally incapacitated and that we are capable of ceasing altogether to engage or entertain reactive attitudes insofar as we have reason to believe that *everyone* is incapacitated in the relevant ways. If the thesis of determinism is true, the Pessimist argues, then we are, indeed, all morally incapacitated. (Russell 1992, 296)

On the one hand, I am compelled to agree with both Nagel's objections to and Russell's worries about Strawson. We are capable of assuming the objective view and thus suspending reactive attitudes toward all agents. But, on the other hand, as

---

[58] Russell's paper is much more detailed in its arguments and the paraphrase I propose to offer may not be fully reflective of Russell's arguments, but as far as I understand him and for the purposes of this essay, the paraphrase will be enough to highlight Russell's point and illustrate a problem for Strawson.

evident in Nagel's "Lieutenant Calley" example and in thinkers like McDermott, the subjective view (or the experience of freedom) is inevitable. In other words, we are bound to slip back into the inner (subjective) view. The reality of the subjective view (the experience of the world from the inner standpoint) makes it impossible for us to consider the actions of agents from an outer view. If reactive attitudes are inseparable from the subjective view, and if the subjective view inevitably resurfaces in certain circumstances or under certain conditions (circumstances and conditions defined by our everyday lives and thus as being distinct from our philosophical ponderings and meditations), then, at the time we operate from within the subjective standpoint, we unavoidably must work with the framework of reactive attitudes, which resurfaces along with the inner view.

Thus, even though Strawson may be wrong in claiming that we are unable to suspend reactive attitudes toward all people, his position is correct in that we are unable to entertain the objective view indefinitely (in fact, we are unable to retain such a view for any considerable or extended period of time). Therefore, reactive attitudes are as certain to resurface or re-emerge as our subjective view is and, I am compelled to argue, are as inevitable as phenomenal consciousness is. In other words, we are hardwired to be disposed to (and in the right circumstances to act on) our reactive attitudes.

Consequently, not only are we hardwired to feel free (due to our subjective or inner viewpoint), but we are also hardwired to respond reactively (as if we were free) even given the knowledge that the mind is merely mechanistic or that the thesis of determinism is true.

If McDermott's notion of the self-model is an accurate approximation to (or a good analogy for) the structure of the human mind and the mind's perception of self, then freedom of the will is an illusion; agents and their actions fade into "the flux of events in the world of which the agent is a part" (Nagel 1986, 229). In other words, if McDermott is correct, the equivalent of the thesis of determinism is true (even if determinism is not the case, a mechanistic view of mind, dependent on nature and nurture, that is, on inner and environmental causes, amounts to the same type of dissolution of the Libertarian notion of freedom as the truth of the thesis of determinism would).

The Strawsonian notion of reactive attitudes appears to offer, in such a case, a good explanation for why human beings are moral creatures and it seems to provide us with the type of "freedom"[59] we crave, namely one that can justify the practice of praising and blaming.

3.6
Hardwired Freedom and Responsibility

As already argued, I do not read McDermott's vision of freedom as a self-fulfilling illusion. The reason for this is that the kind of freedom McDermott postulates is merely an experience of freedom arising out of the nature of the self-model. The fact that our deliberative mechanisms are transparent to us, that we are incapable of abandoning the subjective point of view for very long due to the properties of mineness, selfhood, and perspecitvalness (which arise in the self-model), that we are unable to completely model ourselves (predict ourselves), and the fact that

[59] Although as already argued above, metaphysical freedom is not necessary for the assignment of responsibility, praise, blame, etc.

we endorse our actions by, for instance, confabulating the necessary intentions all contribute to the illusion of freedom.

Our self-models are incapable of abandoning this illusion of free will. At most, our philosophical ponderings allow a glimpse into the objective realm via the adoption of the objective standpoint. The feeling of freedom is hardwired into our self-models as the reactive attitudes are hardwired into our social nature. And although we can withdraw into the objective point of view, as soon as we fall back into the subjective, participant stance (to which we are bound to return as soon as we resume our regular interaction with the world and other people), the hardwired reactive attitudes and the ever-present feeling of freedom force themselves upon our experience.

I suspect that the commonsense intuition that responsibility requires metaphysical freedom rests on the fact that when we operate from within the subjective standpoint, we feel free. Reactive attitudes, as Smilansky suggests, may in fact require illusion (or to put it differently, the illusion of freedom may in fact be a necessary prerequisite for the stability of our reactive attitudes). That is, because I feel free, I am capable of holding myself accountable for doing A instead of doing B. Without the illusion of freedom, one could perhaps spend much more time in the objective realm (it would, at least, be much easier to do so). But, since as McDermott argues, the illusory feeling of freedom is an inevitable consequence of self-modelling, we are forced into the subjective view every time we succumb to the illusion (and we succumb to it nearly all our waking lives).

If reactive attitudes are intimately connected with the experience of freedom, as I believe they are, then it would appear that ridding ourselves of these reactive attitudes is as impossible as ridding ourselves of the feeling of freedom. Both are the result of self-modelling and both are inescapable realities of our subjective lives.

Responsibility, on such a view, would have to be the product of our models of the social realm. In other words, just as we model ourselves as free and just as our self-models model our bodies, the social space must also be modelled by the systems that interact in it. We may not require metaphysical freedom to be held responsible, but the illusion of freedom seems to be a convenient vehicle for responsibility.

The problem of the arbitrariness of second-order volitions we encountered with Frankfurt's view also holds in the case of illusory freedom. However, responsibility seems to occupy a slightly different niche than the illusion of freedom does. Although responsibility does not enjoy the same degree of reality as a tree or a rock, for instance, it does not merely reside in the self-model (as does the experience of freedom).

Moral accountability is a real social practice. Although, from the objective standpoint, we cannot pick out responsibility in the world because it does not feature there,[60] we do recognize it in the social context (from within the subjective point of view) of which our reactive attitudes are a natural part. The difference between our experience of freedom and our notion of responsibility, then, is that whereas the illusion of freedom is confined to the self-model, our understanding of responsibility exists in virtue of the inter-subjective nature of the social framework. In other words,

[60] This is why every time we adopt the objective standpoint, we recognize the ultimate arbitrariness of passing moral judgments.

responsibility appears to occupy the inter-personal sphere of the social realm while freedom inhabits only the intra-personal space of the self-model. But both responsibility and the experience of freedom are products of the subjective experience (subjective standpoint) arising out of the process of self-modelling.

As Smilansky so aptly observed, the moral house we have is none other than the "Community of Responsibility." Our moral house, then, is the social realm where our reactive attitudes guard, and partly dictate, the social rules that those who are members of the moral household must adhere to and adopt. The subjective standpoint, which is generated by the self-model, perpetuates the illusion of freedom and houses the reactive attitudes. The subjective view, the experience of freedom, and our reactive attitudes are intimately connected; they buttress each other and interact with each other. In effect, they all stem from the nature of our self-models and are as inevitably human as the self-model itself.

The objective point of view, though accessible, is not a stable feature of human experience. Having said this, however, it casts a dreadful shadow on our humanity. The objective view reveals our insignificance, helplessness, lack of control, and hints at the fundamental arbitrariness of our actions, judgments, beliefs, and practices. However, even if the social sphere, the realm where the concept of responsibility is meaningful, supervenes on the objective reality of the physical world, the ultimate arbitrariness and fundamental injustice of our moral judgments and practices becomes invisible within the context of the social.

We live in both realms (the subjective and the objective), but we are moral agents only within the subjective sphere. McDermott's physicalism, to the best of my

understanding, not only does not deny such a conclusion, but actually contributes to it. The self-model weaves a fictitious world for itself and lives in it. And although neither freedom nor moral responsibility exist objectively (but only within this fiction), human beings can be held responsible even if they are not ultimately free because whereas the freedom we believe we have does not exist (since we believe we are objectively free, which, according to my understanding of McDermott's view, we are not), the responsibility we believe we have does exist (since we believe responsibility to be a feature of our social world, our "Community of Responsibility," of which it, in fact, is a feature). The Community of Responsibility exists in virtue of the fact that we are social beings and that we engage in inter-personal relationships, which are, in fact, objectively real. Responsibility is relevant only within the context of these inter-personal relationships. There is no other sense of responsibility or morality. This is why we do not hold trees, rocks, tigers, or household items morally accountable. In fact, as already mentioned, we do not include infants or the insane in the category to which the concept of responsibility is applied. Thus, whereas free will requires grounding in objective reality in order to be more than an illusion, it does not even make sense to demand that responsibility be grounded within anything other than the subjective realm, which in turn is based in the objective reality of inter-personal relationships. But, the illusion of freedom is necessary insofar as it perpetuates the reactive attitudes.

Chapter Four
Looking for Freedom in Other Places[61]

4.1
On Second Thought...

Searle states that the reason many philosophers are drawn to compatibilism is that "they are not really very much interested in the problem of free will...They are interested in the problem of "moral responsibility"[62] (Searle 2004, 156). Searle, on the other hand, is concerned with the problem of what van Inwagen refers to as metaphysical freedom.[63] Searle assumes, for the purpose of his argument, that psychological freedom is real. That is, he assumes that our psychological states are not causally sufficient to determine all of our voluntary actions.

The deep question for Searle concerns the underlying neurobiology. Searle states his worry as follows:

> We might have free will at the psychological level in the sense that the psychology as such was not sufficient to fix our actions. But the underlying neurobiology, which also determines that psychology, might itself be causally sufficient to determine our actions. (Searle 2004, 158)

In other words, if the psychological is nothing more than the neurobiological described at a higher level, the psychological freedom (the gap that is characterized by the lack of causally sufficient psychological conditions to determine our voluntary

---

[61] Although as far as McDermott is concerned, I have my doubts that free will, on his view, could be construed in the libertarian fashion, I think that the free will debate is far from being resolved. And so, I wish to engage in an exploratory project in this chapter, which is not meant to be conclusive in any way, but rather is meant to open the doors to the possibility of libertarian free will. I expect that this largely incomplete contemplation will prove to be the beginning of future musings. And so, even though this chapter will be mostly speculative, and even if I shall mostly bracket the previous chapters, I feel the need, given the free will discussions of the first three chapters, to at least flag the possibility I entertain here.

[62] This is, in fact, the route I decided to take when analyzing McDermott's view.

[63] See van Inwagen, P. (1998). "The Mystery of Metaphysical Freedom." *Metaphysics: The Big Questions*. Malden, MA: Blackwell Publishing Ltd. 1998. 365-374.

actions) must go all the way down to the neurobiological level. "But," Searle writes, "how could it? There are no gaps in the brain" (Searle 2004, 159).

If we adopt what Searle calls the Mechanical Brain Hypothesis,[64] then even if the mind were given the illusion of free will,[65] the whole system would remain deterministic. In order to arrive at the notion of metaphysical freedom, according to Searle, we must appeal to the only source of indeterminism presently known to us in nature, namely quantum indeterminism. And so, the brain, on such a view, would have to be a Quantum Brain.

Whereas the Mechanical Brain Hypothesis makes the claim that the state of the system at t_1 is causally responsible for the state of the system at t_2, the Quantum Brain Hypothesis makes the claim that at the quantum mechanical level, the state of the system at t_1 is only causally responsible for the state of the system at t_2 in a statistical manner because there is a random element at work at the quantum mechanical level. A common argument, one that Searle admittedly found convincing in the past, is that quantum mechanics gives us randomness, but not freedom.

Recently, Searle has argued that such an argument commits the fallacy of composition. "The fallacy of composition is the fallacy of arguing from properties of the parts of a system to the whole system" (Searle 2004, 162). David Cole, in his

---

[64] The Mechanical Brain Hypothesis is, in fact, the way I interpret McDermott's account and thus, as already argued, I think of McDermott's view in terms of an illusionist account of freedom. However, even though I do not read McDermott's account of mind in terms of the Quantum Brain Hypothesis, I do not have much trouble imagining that the brain may, in fact, turn out to be a quantum system (though I definitely do find it quite difficult to imagine and understand what precisely that would mean or entail).

[65] Searle does not think that the Mechanical Brain Hypothesis necessarily runs counter to our experience because, after all, we have all kinds of illusory experiences.

article "Thought and Thought Experiments,"[66] gives an example of the fallacy of composition. He writes:

> Imagine a drop of water expanded in size until each molecule is the size of a grindstone in a mill.[67] If you walked through such a now mill-sized drop of water, you might see wondrous things but you would see nothing *wet*. But this hardly shows that water does not consist *solely* of H_2O molecules. Rather, it shows that a fallacy of composition is at work here...it shows that whenever one takes the *perspective* of the subsystem or constituent, one is likely to find it hard to believe that the whole and all its properties can be accounted for by the properties of the constituents, given their arrangements. And this difficulty increases with increases in the complexity of the systems and their global behavior. (Cole 1984, 432)

What I understand Searle's insight regarding the fallacy of composition to imply is that our Self Forming Actions[68] are in fact instances of free will at work and not merely random events. However, we must ask ourselves what sort of process or structure may be responsible for shielding the higher-level phenomena from the randomness inherent at the lower-level?

4.2
Choice and Indeterminism According to Nozick

Robert Nozick, in Chapter Four of *Philosophical Explanations*, outlines an indeterministic decision-making framework that is similar to that presented by Robert Kane (whose terminology of 'Self Forming Actions' I am employing here). On

---

[66] David Cole attacks Searle's Chinese Room Argument in this article.

[67] Cole is applying the fallacy of composition to Leibniz's comparison of the mind to a mill.

[68] When making use of this term, I have in mind Robert Kane's libertarianism. The centrepiece of Kane's theory is his notion of Self-Forming Actions (SFAs). According to Kane, since the agent is responsible for the formation of her character (via SFAs), she must also be responsible for the actions that are a direct result of her character. Kane explains that SFAs occur at times in life when we are torn between competing visions of what we should do or become. What happens during such moments of inner conflict is that a tension and uncertainty arises in our minds, "a kind of stirring up of chaos in the brain that makes it sensitive to micro-indeterminacies at the neuronal level" (Kane 1999, 306). The inner conflict originates in the desire to both do *A* and *B* (where doing *A* is incompatible with doing *B*). Overcoming temptation, then, is the result of our effort while failure to overcome temptation comes about due to the fact that we did not allow our effort to succeed (this is owing to the fact that we wanted both to overcome the temptation and to fail). However, whatever choice we make, it will be our own choice because we were trying to make both *A* and *B* obtain. According to Kane, then, there is room for shaping our own characters on such a view. In terms of McDermott's robot (just to tie it back to McDermott), we can imagine that R is trying to perform G1 and G2 simultaneously (because the values assigned to each goal are the same and thus, its decision-making system has no way of deciding for or against a certain goal). Both G1 and G2 are assigned the same preference status.

Nozick's view, in short, human beings have the power to form their own characters by having reasons for their actions and by assigning weights to these reasons. (This strikes me as being quite similar to McDermott's notion that Rosie the robot labels her goals as preferences). Self Forming Actions[69] occur when the assignment of weights to reasons contributes to the formation or reformation of oneself.

Nozick claims that free actions are caused, but undetermined. To paraphrase Nozick, let us imagine that whether an agent *S* does *A* or *B* is undetermined (*S* could do *A* in circumstance *C* while in mental state *M* and *S* could do *B* in identical circumstance *C* while in identical mental state *M*—whether *S* does *A* or *B* is genuinely indeterminate). But if *S* does *A*, then her doing *A* is caused by her reasons R_A. If *S* does *B*, on the other hand, then her doing *B* is caused by reasons R_B.

Also, *S* both weighs and weights reasons. That is, *S* assigns weights to reasons (*S* weights them) and *S* weighs the weighted reasons according to the weightings *S* assigns. The weightings, to the best of my understanding, are indeterminate (this is why *S* can either choose to go along with R_A, which will be the cause of *A*, or R_B, which will be the cause of *B*). The weights need not be exact as long as some reasons R_x outweigh other reasons R_y because any inequality in weight will be sufficient for choosing, say, R_x over R_y and thereby causing an action *X* rather than *Y*.

Nozick considers an indeterministic self-determining decision-making process. He makes the analogy with the currently orthodox quantum mechanical

[69] Nozick does not call them SFAs, but I shall make use of this terminology as there are similarities between Nozick's and Kane's view.

theory of measurement,[70] which states that a quantum mechanical system is in a superposition of states, which changes continuously in accord with the quantum mechanical equations of motion. A particle in superposition can be imagined as possessing multiple positions or states simultaneously. According to the orthodox interpretation of quantum mechanics, a superposed particle collapses into a single state once measured or observed. Such collapses are probability-driven in that it is genuinely unpredictable which of the numerous simultaneously existing states will actually obtain.

Nozick draws an analogy between the effect observation/measurement has, in quantum mechanics, on the wave packet and the way in which indeterministic decision-making processes may be responsible for freedom of the will. Nozick writes:

> [A] person before decision has reasons without fixed weights; he is in a superposition of (precise) weights...The process of decision reduces the superposition to one state (or to a set of states corresponding to a comparative ranking of reasons), but it is not predictable or determined to which state of the weights the decision (analogous to a measurement) will reduce the superposition. (Nozick 1981, 298)

Nozick is not endorsing the orthodox account of measurement in quantum mechanics as correct, but only draws upon its theoretical structure to prove that his indeterministic conception of decision is a coherent one. "Decision fixes the weights of reasons; it reduces the previously obtaining mixed state or superposition. However [and this is the crucial point], it does not do so at random" (Nozick 1981, 299).

Moreover, according to Nozick, these bestowed weights (or comparative weightings of reasons) do not disappear immediately following the decision that

[70] Professor Wesley Cooper, in personal correspondence, pointed out to me that perhaps this is no longer the currently orthodox view especially considering the growing number of proponents of the many worlds interpretation of quantum mechanics. I shall, however, concern myself with the kind of collapse theory made use of by Nozick in his contemplations.

bestows them. These weightings "set up a framework within which we make future decisions" (Nozick 1981, 297). Nozick compares this framework to precedents within a legal system. That is, just like precedents in legal cases, "the decision represents a tentative commitment to make future decisions in accordance with the weights it establishes" (Nozick 1981, 297). Nozick makes it clear that we do not always need to act on such precedents (even if they were, at the time of being set, the strongest preferences or motives). A given reason can always "become strongest in the process of making the decision, thereafter having greater weight (in other future decisions) than the reasons it vanquished" (Nozick 1981, 297).

The problem to be resolved, however, can be phrased as follows: how are such weightings not random? In other words:

> [H]ow does Nozick reply to the claim that it is an arbitrary, random matter what character the reasons-weighting decision will have, and so it is not in any clear sense up to the agent what the decision will be? (O'Connor 1993, 512)

Nozick's reply goes as follows: if *S* decides to adopt the policy of tracking bestness, for instance, then the decision to do so is itself in accordance with that very policy. Similarly, if *S* assigns weights to reasons and bases such assignments on a policy of tracking a previously chosen conception of oneself and one's life (I would imagine that such a previously chosen conception of oneself is chosen via the self-forming decisions one may have been faced with in one's past), then the weightings one bestows will result in a self-subsuming decision because the conception of herself, which *S* chose to track includes bestowing those very weights (or at least similar weights) and choosing that very conception.[71]

[71] However, as will become evident in the following section, self-subsuming decisions, in and of themselves, do not solve the problem.

4.3
The Problem of Arbitrariness

It would prove useful to review Robert Kane's view of Self Forming Actions in light of Nozick's notion of an indeterministic decision-making framework. Let us fill in the necessary details. I shall make use of Kane's example of a businesswoman torn between two choices. The businesswoman, call her Sally, is on her way to a meeting when she encounters a mugging. She can rush to the bus stop and make the meeting at the price of letting the victim fend for him/herself or she can call for help at the cost of missing her (extremely important) meeting. The question, then, is: what precisely is going on during her decision-making process, which let us imagine, terminates in the noble act of aiding the victim?

Kane proposes that Sally wants both options to be actualized (she desires both to make her meeting and to help the victim). We can label the first option as *A* and the latter as *B*. What is indeterminate in this case is not the question of which options are available to Sally (because we can imagine that Sally's character, through previous, precedent-setting Self Forming Actions, is such that Sally is bound to want both *A* and *B* to obtain), but which possible course (*A* or *B*, but not both) she will decide to pursue.

Kane differs from Nozick in that on Kane's view "[i]nstead of the weight of reasons being actually indeterminate prior to choice, they have values at any given time, but unstably so, fluctuating in an undetermined manner until the moment of choice" (O'Connor 1993, 520) whereas for Nozick, the weightings are being set at the time of the decision-making process. Kane, however, does share with Nozick the

notion that "the choice brings it about that one set of reasons prevails, and issues in the corresponding action" (O'Connor 1993, 520). Ignoring the difference between Kane and Nozick,[72] however, Sally's choosing to pursue *B* rather than *A* is a result of fixing the weightings such that the reasons Sally has for *B* are stronger than those she has for *A*. And since the assignment of weights to reasons is a key feature of Self Forming Actions, it must be here that the relevant indeterminacy must be shown to make such assignments undetermined, but non-random.

Because Sally wants both *A* and *B*, even a random choice will necessarily result in a desired outcome. The question that concerns me here is what makes Sally's choosing *B* over *A* such that it is actually up to Sally to choose *B*. That is, can the indeterministic decision-making process(es) she employs in making her choice actually account for free will (metaphysical freedom)?

It would appear as though self-subsuming indeterministic decisions are autonomous enough to ensure not only that alternative possibilities are open to the agent, but also that the agent has enough of the relevant kind of control necessary for choosing between these possibilities. The question, however, remains whether it is not arbitrary "that one self-subsuming decision is made rather than another" (Nozick

[72] Since for Kane, the weightings are already set, the indeterminacy may actually work out slightly differently than for Nozick on whose view the weightings are being set at the time of the decision-making process. However, one could imagine that the indeterministically fluctuating weightings on Kane's account are analogous to the cases where, on Nozick's view, the agent is in the process of fixing weights, but where the fixing of weights occurs against the background of previously set precedents (which I would imagine is the case for most if not all actions, given that most agents/persons whom we would consider holding responsible for their acts have lived through enough self-forming moments as to have a robust framework of precedents upon which to draw). And so, it may not be of great importance whether we view Sally as indeterministically fixing weightings or indeterministically choosing between already weighted possibilities (where the weightings are themselves prone to fluctuation).

1981, 301)? That is, will it not prove impossible to explain or account for why self-subsuming decision *A* was made rather than self-subsuming decision *B*?

Nozick argues that free decisions are not produced by a random chance mechanism because the process of choice among alternative actions is different. There are no fixed factual probabilities for each action, but, rather, there is a process that makes it possible for a number of actions to occur where one of them actually does. One can say that "[t]his time, the process gave rise to that particular alternative. (Compare: this time the random system yielded that particular event)" (Nozick 1981, 302).

I imagine that Kane, Searle, and Nozick all share the same intuition, namely that if quantum indeterminism is the key to unlocking the mystery of metaphysical freedom, then whatever processes are involved, they must be indeterministic, but non-random. Perhaps Nozick is on to something when he states: "at least we may see the later adherence to weights as an indication of their non-random character; if the choice of these weights was simply random and arbitrary, would they win continued adherence" (Nozick 1981, 306)?

I think the intuition is correct. It seems quite unlikely that if each decision were ultimately governed by random events, one could ever be truly set in one's ways (as we often observe in people with substantial life experience). But, perhaps continued adherence is not an indeterministic, but a deterministic feature that depends on the process(es) responsible for making use of the precedents set during moments of moral tension (during our Self Forming Actions). The initial problem, it seems, still remains to be resolved.

Presently, I would like to consider and work through O'Connor's objection to Nozick. Later, as a means of defending Nozick's view from the randomness objection raised by O'Connor, I will consider the role experience and the setting of precedents may play in minimizing the effect of randomness and, in fact, actually utilizing indeterminacy in the complex process of making freely-willed decisions.

O'Connor phrases the problem in the following manner:

> Suppose we have a decision in which the weights it bestows "fix general principles that mandate not only the relevant act but also the bestowing of those (or similar) weights." Can we explain how it is up to the agent that those weights (rather than some others) are assigned, by noting that the decision is "an instance of the very conception and weights chosen"? (O'Connor 1993, 514)

O'Connor states that he cannot see how we can. He claims that there is a "disanalogy between an explanatory law and a decision that *institutes* a general policy or conception in accordance with which one is to act" (O'Connor 1993, 514). He continues:

> Explanatory laws do not become true at some moment in time. If there were an analogy here, it would have to be between self-subsumptive decisions and an event of a law's coming to be true, which event was explained by, because subsumed under, the very content of the law. (O'Connor 1993, 514)

In other words, the disanalogy argument put forth by O'Connor, as I understand it, appeals to the intuition that Nozick needs to explain why we cannot say that choosing a particular conception of oneself (or tracking bestness) is an arbitrary or random choice. That is, O'Connor argues that if there were an analogy, it would be between self-subsuming decisions and an event (a particular instance) of a law's coming to be true where the event would be explained by the content of the law itself because it (the event of a law's coming to be true) would be subsumed under the content of the law. As I understand it, the problem is that the event may be justified, but the law itself would need to be timeless (and universal) for the event to be subsumed under its

content. And if this analogy (between self-subsuming decisions and an event of a law's coming to be true) is to hold, then a self-subsuming decision requires a general policy such as tracking bestness or a previously chosen conception of oneself etc., which would be the analogue of the law. So far so good, but since such a conception of oneself or one's proneness to adopt a policy of tracking bestness is not timeless nor is it universal (but needs to somehow non-deterministically and non-randomly be adopted), the actual adoption of such a general policy (i.e. tracking bestness or tracking a previously chosen conception of oneself) will either be determined from the outset (perhaps genetically pre-programmed and nourished by certain environmental contingencies, or in some other way) or will be indeterminate. But if the adoption of a certain general policy, which seems crucial for self-subsuming decisions, is in fact indeterminate (or, alternatively, determined but out of the agent's control), then one is once again faced with the prospect of arbitrariness since the decision to select a certain general policy is itself not a self-subsuming decision. And so, the problem does not seem to go away.

O'Connor states the above argument much more eloquently and thus, I shall cite it here. He writes:

> If I come to order certain values…during a decision-making process, this ordering subsequent to the decision will "affirm" that decision precisely because the latter is an act that is in accordance with it…[and in a footnote, he continues,] a person's subsequent affirmation of a decision that marked a restructuring of values may come from those priorities having become deeply entrenched over time, whereas the initial adoption of them was somewhat tentative and experimental. (O'Connor 1993, 514-515)

The initial adoption of the general policy had to have been somewhat tentative and experimental because there would not have existed any prior reasons or general policies under which the initial decision (to follow a certain general policy) could be subsumed. And if indeterminacy is necessary to make a free decision non-

determined, but is not sufficient to make it non-random, then without self-subsumption (which is, according to Nozick, sufficient to make a decision a non-random one), all we appear to be left with is the arbitrariness of quantum indeterminacy (at least during the initial formative stages where we adopt certain general policies for the very first time).

At least at first blush, it would appear that if we allow for randomness at the formative stages, then even if the consequent decisions are non-random (because they are self-subsuming), they are ultimately governed by a capricious quantum hiccup and thus the people we know ourselves to be (i.e. I "know" I am a person that would most probably favour decision *X* over decision *Y* because I "know" that, at least in most circumstances, I would "choose" to follow a general policy *Px* over another general policy *Py*) and the selves we experience as making controlled, rational, and free choices are the outcomes of a cosmic coin-toss. This is, I think, a fair *prima facie* intuition to hold. It is an intuition John Searle admittedly adhered to for a number of years. But, of course, this is also the very same intuition Searle questions in his 2004 book *Mind: A Brief Introduction.*

There is an enormous difference between Sally's decision-making process and the probabilistic nature of radioactive decay. Both Sally and a radium atom, at the most fundamental level, are governed by quantum indeterminacy, but unlike a radium atom, Sally's decisions consist of a complex system of beliefs, desires, motivations, a rich history of past decisions, and a continually evolving self-conception, etc. Although the probabilistic (or random) effects of quantum indeterminacy do play an important role in Sally's decision-making processes, Sally, throughout her entire life,

acquires various motives for her actions and has certain dispositions governing her future choices.

It is important to draw a distinction between an experienced Sally and an inexperienced (callow) Sally.[73] Callow Sally, say Sally at age 20, faced with the difficult choice between stopping to aid the victim and making it to her important meeting may be much more susceptible to the probabilistic quantum effects than the experienced Sally, Sally at age 40, for instance.

I would like to revisit Kane's example with the appropriate alterations. Let us imagine that callow Sally, having spent most of her life in a small quiet town (a community where no one feels the need to lock their doors at night), moves to the big city in search of work. Callow Sally, on her way to her very first job interview at an established firm, encounters a mugging. She knows that she will not make her meeting unless she runs to the bus stop, but, having had the privilege of a well rounded moral education, she weighs the two decisions (or assigns weightings to each in accord with some set preferences) and proceeds to make her choice, which, as it turns out, amounts to the noble act of helping the victim. At the time of the choice, however, Sally is struggling with this difficult decision. She wants both to help the victim and to make it to her meeting. Callow Sally's choice to provide aid at the price of missing her meeting is a difficult one, one that could have resulted in the alternative course of action where Sally goes to the meeting instead. In other words, Sally could have done otherwise (she could have chosen to go to the meeting), but she opted for the noble alternative and it was far from certain which choice she would commit herself to because she wanted both (to go to the meeting and to help the

[73] I would like to thank Professor Wesley Cooper for pointing out this important distinction.

victim) and both options were open to her at the time of the choice. And, being largely inexperienced in making such decisions, her choice, though noble, was dependent partly on her upbringing, partly on her self-conception (what kind of individual she sees herself as), and partly on randomness (even callow Sally herself could not have predicted what choice she would be likely to make since she has not had much time to think about the possibility of being faced with such problems and she has not had the relevant life experience, which might contribute to making one course of action more likely than the other).

Sally's choice has various consequences one of which is the fact that her phone call to the police saves a life, but another being the fact that she does not get the job she wants. Let us imagine, however, that Sally ends up finding another equally good job and spends the next twenty years becoming very successful. The now experienced Sally (having had made numerous moral decisions and having experienced a variety of different situations under a number of different circumstances including the decision of twenty years ago, which had a considerable impact on her life) is hurrying to a very important meeting with very influential clients. Her career depends on the outcome of this meeting. She manages to hail a cab and starts heading in the cab's general direction while the impatient driver keeps honking his horn in order to get her to hurry because he has many other clients waiting since taxis are hard to find at this busy time of day. As the experienced Sally runs toward the cab, she notices that the friendly storeowner (who is also her neighbour) is being robbed at gunpoint. She reaches for her cell phone while running toward the cab so that she can both catch the cab and make her meeting and notify the

police about the robbery in progress and thereby aid her neighbour. Unfortunately, in her haste and anxiety, she drops her cell phone and it shatters on the paved sidewalk.

The experienced Sally knows that if she goes to make a phone call from the public phone booth, which she just passed, the impatient taxi driver will not wait for her and she will not be able to catch another taxi in time to make it to her meeting. On the other hand, she also know that if she gets into the cab, her neighbour's store will be robbed and he may be seriously harmed. Although the experienced Sally finds herself in a similarly difficult situation as that of twenty years ago where the stakes are equally high and even though her decision is undetermined because both courses of action are open to her and she wants to pursue both, but can only decide to act upon one of them, the experienced Sally (having spent the last twenty years and the twenty years before that tracking bestness and tracking a conception of herself that is compatible with tracking bestness) is less likely to choose randomly. Her preferences are much more solidly set than the preferences of her younger self. And, given her character and the numerous moral precedents set and reinforced by past actions, the experienced Sally, after a moment's hesitation as she re-evaluates (weights) her preferences, turns to the phone booth and makes the call while the cab driver sounds his horn one last time and gestures toward the phone booth while he speeds away.

The process of weighting preferences, it could be imagined, is very speedy for the experienced Sally since her preference to aid those in need is buttressed by a long, rich history of choosing selflessly and thus, her weighting of preferences takes into consideration the numerous precedents she has set by means of previous decisions.

The indeterminacy (the openness of both courses of action) is much more definitely closed or narrowed to the one course of action she is most likely to take because, it could be imagined, her preference to aid her neighbour greatly outweighs her preference to make the meeting whereas callow Sally twenty years ago did not have the benefit of such a long history of precedents to contribute to the choice (this is why callow Sally's noble action was much less certain than the experienced Sally's noble act).

The difference between callow Sally and experienced Sally, then, is that the experienced Sally has a more robust framework to bring to the current choice situation than did her younger self. And thus, callow Sally is much more vulnerable to indeterminacy than her more experienced future self. That is, callow Sally is much more vulnerable to moral luck as she makes the noble choice twenty years ago than the experienced Sally whose moral framework is made more stable by the moral precedents she set through the experience of making moral choices. This also means that alternate possibilities open to experienced Sally may be quite narrow (amounting, perhaps, to being able to choose to do the noble thing in slightly different ways, but where doing the selfish thing is not an option). The experienced Sally, however, does not feel as though she is limited in her freedom because she is acting in accord with her character (she is being herself and thus is acting in accord with her will).

4.4
Finite Beings, Finite Freedom?

Searle writes:

> It may well be that the evolutionary function of consciousness is at least in part to organize the brain in such a way that conscious decision making can proceed in the absence of causally

> sufficient conditions even though the effect of conscious rationality is precisely such as to avoid random decision making. (Searle 2004, 162)

It seems to me that one innovative attempt at outlining the kind of conscious decision-making process that proceeds in the absence of causally sufficient conditions and yet avoids randomness to which Searle vaguely alludes is that described by Nozick (as sketched out above by the example of callow and experienced Sally).

Although many thinkers find it difficult to abandon the intuition that any antecedent randomness will inevitably pierce through any amount of complexity only to resurface at the higher level of a system, I think that the difficulty stems from the fact that they lack a clear picture of what kinds of processes could be responsible for the absorption of randomness and a consequent metamorphosis of arbitrariness into free will. Nozick's innovative account seems to be a good attempt at outlining some such process.

There is, I think, a problem with criticisms of the nature given by O'Connor. First, I take it that everyone will agree that human beings are finite creatures. That is, we have a beginning and an end. Presently, let us just consider the beginning. Any talk about freedom of the will when applied to human beings should, I would imagine, be concerned with living, embodied members of the species homo sapiens. That is, when we are talking about free will, we are talking about existing, physical agents. Finite creatures (organisms that come into existence sometime after the big bang) that exist in the physical universe lack the God-like quality of self-creation. In other words, we do not have control over how, where, when, to whom, with what genetic material, etc. we are born. There is a certain self that is simply given. This, in part, is what it means to be finite: we *begin*, but cannot *come to be* out of nothing

and thus are dependent on whatever exists prior to our beginning. And, despite such humbling beginnings, we rightly identify ourselves with what is simply given: I am the continuation of the body that began at conception, I am that and much, much more: I am my genetic material, my brain, my mind, etc. All of these are simply given, in some form, right from the start. And this initial givenness is what I am. Whether or not I am free may not necessarily have anything to do with whether or not I have the divine power to will myself into existence (which obviously no finite being can have).

Human beings are born with quite impressive hardware and some amount of state-of-the-art software ready to go with an amazingly user-friendly interface (which may develop into something even more remarkable—i.e. a self-reflexive and conscious self-model—with time). In short, we come fully loaded and with some amount of upgradeable potential.

Our formative years are also given in a sense because so many aspects of our personalities will ultimately depend, in large part, on luck. That is, we are shaped into "us" by: the people that raise us, the genetic makeup of our biological parents, by individuals and groups we interact with, by the social and cultural conventions of our surroundings, as well as, if the indeterminist libertarian is correct, some amount of randomness. Callow Sally's choices will be mostly grounded in such givenness and due to her inexperience as well as a lack of a robust enough moral framework, which stems from the fact that her character is very sensitive to outward pressures and forces, she will be more vulnerable to and more swayable by the effects of randomness inherent in quantum indeterminacy.

At some point, however, we begin to make choices in accordance with the general policies handed down to us by our parents, society, etc. It is not really our fault nor can we be credited with the possession of certain general policies (i.e. tracking bestness), but we are certainly praised and blamed in accordance with how well we follow these cultural, societal, etc. policies. This is why I feel that responsibility is to be found in the social realm. With time, as was the case for experienced Sally, even though the underlying indeterminacy ensures the openness of the future, our more mature choices are no longer affected by the indeterminacy in a whimsical manner as was the case for our younger, less experienced selves. And although the experienced Sally's character is less malleable, it remains shapeable, but in contrast with callow Sally, is also much more stable and predictable.

And so, perhaps we could simply view the general policies handed down to us as being part of our given nature, which we appropriate with such ease. I must admit that the picture presented so far (that much of what defines us is simply given) has a heavily deterministic feel to it. And I think that is as it should be because I believe that much of what defines us is ultimately out of our control. However, the question I am presently concerned with is whether we have free will and not whether we have a self-creative power. If self-creation is necessary for freedom of the will, then we might as well end the discussion here and now. It seems to me that what thinkers like Nozick and Searle are after is an explanation of how we can make free decisions given the natures we have.

Since human beings are finite, why can we not say that freedom of the will is also finite? I do not think that it makes sense to suppose that newborn babies exercise

free will. Perhaps free will also has a beginning (like we have a beginning), but freedom of the will only begins (emerges) in the right context (a certain amount of complexity and within certain psychological and physiological structures). If the above is true, then perhaps we have free will. But such freedom is not our birthright. Rather, it is a freedom we grow into and must labour for. Not only are there many constraints on such freedom, but also, it comes in varying degrees.

If freedom is basically self-formation (and not self-creation), then there may be room enough for Nozick's account after all. If we take the initial (even randomly selected) general policies (i.e. of tracking bestness or tracking a previously chosen conception of oneself, etc.) as simply given, then self-subsuming decision-making processes (which are subsumed under the "given" general policies) can still be quite powerful self-forming and self-reforming tools. In other words, if we are given a certain material (i.e. our genes, talents, some dispositions, certain beliefs, etc.), we are powerless when it comes to changing anything, but there may be room for shaping the material already given. Maybe just as heat makes metal more malleable, a healthy dose of randomness can soften up the *given material* to make it shapeable. And this, it would appear, is where Nozick's model fits in.

It is by no means clear whether we should call this free will and whether randomness and causal determination do not actually kill the slightest flicker of freedom. If what I have described above does not amount to free will, then what we experience as freedom is merely an illusion (but this, as already argued, does not take away from our notion of responsibility). And so, even though the compatibility of

quantum indeterminacy with free will may very well be a live option, the debate is by no means near resolution.

In fact, Searle himself is doubtful that quantum indeterminacy can offer anything but randomness. He writes:

> [T]he hypothesis that the random indeterminacy at the quantum level leads to an indeterminacy of a non-random kind at the conscious intentionalistic level, seems very unlikely and implausible. If we are given a choice between Hypothesis 1 [the Mechanical Brain Hypothesis] and Hypothesis 2 [the Quantum Brain Hypothesis], but also given all that we know about nature, Hypothesis 1 seems much more plausible. (Searle 2004, 162)

I believe that thinkers like Searle and Nozick take the issue of free will quite seriously. In fact, they take it seriously enough that both are very careful not to appear as though they are offering solutions, but that they are merely scouting the territory. And, although both feel that quantum mechanics offers a chance of discovering freedom, they are also aware of the fact that more work needs to be done both in quantum physics and on the problem of free will.

Searle invokes the fallacy of composition partly to ensure that the question regarding the possibility of the compatibility of quantum indeterminacy and free will is not briskly abandoned. I think that we should not discount the possibility of "quantum free will." I am quite willing to admit that an attempt akin to Nozick's (as presented above) may in fact succeed in explaining how free will is possible. However, I find myself in agreement with Searle that "we will need to know a great deal more about brain operations before we have a solution to the problem of free will that we can be at all confident is right" (Searle 2004, 151). It is important to note that the existence of free will does not only depend on physics. Whether the world is totally determined, fully random, or perfectly suited for freedom of the will is perhaps quite relevant to the question of freedom, but free will, free choices, and agency are

impossible without the right kind of mind. Having said this, however, if one is a physicalist, the relationship between physics (along with all the laws of nature) and freedom of the will (along with all the necessary complexity of the brain, etc. conducive to freedom) is quite intimate and should not be ignored.

Both thinkers (Searle and Nozick), I think, can be characterized by their inquisitive, but humble approach to philosophical problems. They do not shy away from difficult problems (even the seemingly impossible ones like the problem of free will), and neither of them is afraid to admit that their views may, in fact, be wrong. Searle writes: "We really do not know how free will exists in the brain, if it exists at all" (Searle 2004, 164). And yet, their curiosity seems to always get the better of them. As Nozick confesses: "Over the years I have spent more time thinking about the problem of free will—it felt like banging my head against it—than about any other philosophical topic except perhaps the foundations of ethics" (Nozick 1981, 293).

4.5
Some Benefits of Quantum Randomness: Alms for the Freeless

It is interesting and natural to wonder where the prospect of an incompatibility of quantum mechanics with free will would leave us. One must be cautious not to dismiss quantum indeterminacy as a suitable medium for free will because it may very well turn out that quantum indeterminacy is in fact responsible for genuinely freely willed actions. However, the fact that debates about the compatibility of free will with quantum mechanics are so intense and interpretations of quantum mechanics are so varied is evidence enough to show how difficult the problem of free will really

is and how far from a solution we presently remain. Even if quantum mechanics turns out to be a wrong turn in our adventurous quest for the chalice of dignity (which, for Nozick, the problem of free will is really about), this does not mean that free will is not lurking elsewhere, perhaps never to be discovered, but to be experienced (as it actually is) and to be enjoyed (even if sometimes unwittingly).

If, however, it turns out that quantum mechanics is the last possible site for freedom, and, though rich in many ways, turns out completely lacking in the necessary ingredients for freedom of the will, then what can it offer in return? At the least, if the non-deterministic interpretations of quantum mechanics are correct, then what a person does posses in a quantum mechanical world is an open future.[74] That is, alternative possibilities are still truly presented to us even if we cannot choose which one of the possibilities actually obtains. This is a kind of freedom, although admittedly only a pale, ghost-like shadow of the commonsense, intuitive, and introspectively accessible libertarian vision.

And, if my contemplations of the previous chapter are correct (that is, if the Strawsonian notion of reactive attitudes is rich enough to account for responsibility), then moral responsibility is ours for the taking even in the absence of free will. In other words, if responsibility can truly be separated from the notion of free will, then we are still offered a lesser (perhaps even a considerably much inferior) comfort,

---

[74] Of course, the reality of an open future is argued for by the proponents of the deterministic interpretations of quantum mechanics such as the no-collapse, many-worlds or multiverse view. The future is open, on such views, insofar as there is always a copy of me in a nearby possible world that actually does that which I did not do, but believe that I could have done (if I had not done something else instead). However, since I am presently concerned with Nozick's view, I shall only focus on what Nozick refers to as the orthodox quantum mechanical theory of measurement, I shall not delve into the many-worlds interpretation because such a view raises many problems and questions of its own (i.e. questions about personal identity), which, although immensely interesting, cannot be addressed presently with any amount of detail that would do the topic justice.

namely that of a reality of an open future (which we get from quantum indeterminacy) in conjunction with a sufficiently robust notion of moral responsibility (which, according to Strawson, is independent of whether or not the thesis of determinism is true). Moreover, we can be fairly confident that, at least much of the time, we act in accordance with our characters. Even if on occasion, random wave-function collapses are responsible for the precedents we often appeal to when making choices, a considerable number of our decisions are made in accord with rather stable personalities. Thus, we can still take comfort in the fact that we are on the most part defined by our actions not only because we often act according to our characters, but also because we endorse and appropriate the actions we experience as "freely made" and consequently, form complex and often quite interesting selves.[75]

Another benefit of the reality of quantum indeterminism (even if such indeterminism is incompatible with free will) is that such indeterminacy can serve as an explanatory grounding for our illusion of free will (since quantum indeterminacy can offer a genuinely open future).

I realize that this by no means offers a solution to the problem of free will (or even a solution most people would be comfortable with), but I did not promise a solution. All I intended was the examination of another possibility (one that very well may prove to be a live option). My hope was to get a glimpse of at least a portion of the battlefield on which many more free will disputes will surely take place. Quantum mechanics offers a reasonably new and fertile soil for planting our hopes for

[75] It is true that sometimes we perceive ourselves (and even more often others) as acting randomly (that is, spontaneously and unpredictably) and thus out of character.

free will, but ironically, it is presently unpredictable whether the quantum mechanical realm is capable of hosting autonomous, undetermined, and non-random willings.

Conclusion

I have argued that Peter Strawson's formulation of reactive attitudes solves the problem of moral responsibility with which McDermott's illusionist view of free will is faced and that the notion of reactive attitudes (when viewed as an integral aspect of self-modelling) avoids the awkwardness of acting against our intuitions and in spite of our nature, which may stem from adopting alternative lines of reasoning such as the approach proposed by J. J. C. Smart, which states that we should refrain from making moral judgments altogether.

Moreover, I think that the Strawsonian framework is fortified by McDermott's notion of self-modelling, which grounds the reactive attitudes in the subjective point of view and perpetuates them by means of the hardwired illusion of freedom, which is generated by the self-model.

I have also explored the prospect of the compatibility between quantum indeterminacy and metaphysical freedom. Although I am the first to admit that the issue is still very much an open one, I think that Nozick's approach or something along its lines is very much a live option especially considering that the Strawsonian enrichment of McDermott's view is merely one, but definitely not the only possible solution to the problem of free will and moral responsibility.

Bibliography

Adler, J. E. (1997). "Constrained Belief and the Reactive Attitudes." *Philosophy and Phenomenological Research,* Vol. 57, No. 4. 891-905.

Augustine of Hippo. (391). "Divine Foreknowledge, Evil and the Free Choice of the Will." *Free Will*. Ed. Robert Kane. Malden, MA: Blackwell Publishers. 2002. 259-263.

Brook, A., Raymont, P. (2006). "The Unity of Consciousness." *The Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/consciousness-unity/

Cole, D. (1994). "The Causal Powers of CPUs." *Thinking Computers and Virtual Persons: Essays on the Intentionality of Machines*. Ed. Eric Dietrich. San Diego, CA: Academic Press. 1994. 139-156.

Cole, D. (1984). "Thought and Thought Experiments." *Philosophical Studies* **45**. 431-444.

Copeland, J. (1993). *Artificial Intelligence: A Philosophical Introduction*. Malden, MA: Blackwell Publishers Ltd.

Crane, T. (2003). *The Mechanical Mind: A Philosophical Introduction to Minds, Machines and Mental Representations*. 2nd Ed. New York, NY: Routledge.

Dennett, D. C. (1989). "Evolution, Error, and Intentionality." *The Intentional Stance*. Cambridge, MA: The MIT Press. 287-321.

Dennett, D. C. (1984). *Elbow Room: Varieties of Free Will Worth Wanting*. Cambridge: MIT Press.

Deutsch, David. (1998). *The Fabric of Reality*. London, England. Penguin Books.

Dietrich, E. (1994). "Thinking Computers and the Problem of Intentionality." *Thinking Computers and Virtual Persons: Essays on the Intentionality of Machines*. Ed. Eric Dietrich. San Diego, CA: Academic Press. 1994. 3-36.

Dwyer, S. (2003). "Moral Development and Moral Responsibility." *The Monist Vol. 86 No. 2*. Ed. Barry Smith. Peru, Illinois. 181-199.

Flanagan, O. (2002). *The Problem of the Soul*. New York, NY: Basic Books.

Frankfurt, H. G. (1971). "Freedom of the Will and the Concept of a Person." *Oxford Readings in Philosophy: Free Will, 2nd Ed.* Ed. Gary Watson. Oxford: Oxford University Press. 322-336. 2004.

Frankfurt, H. G. (1969). "Alternate Possibilities and Moral Responsibility." *Oxford Readings in Philosophy: Free Will, 2nd Ed.* Ed. Gary Watson. Oxford: Oxford University Press. 167-176. 2004.

Isaacs, A., Daintith, J., Martin, E. (1999). *A Dictionary of Science*. New York, NY: Oxford University Press.

Jackendoff, R. (1987). *Consciousness and the Computational Mind.* Cambridge, MA: MIT Press. 15-27 (Ch. 2 "The Computational Mind").

Kane, R. (1999). "Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism." *Oxford Readings in Philosophy: Free Will, 2nd Ed.* Ed. Gary Watson. Oxford: Oxford University Press. 299-321. 2004.

Lycan, W. (2006). "Representational Theories of Consciousness." *The Stanford Encyclopedia of Philosophy.* http://plato.stanford.edu/entries/consciousness-representational/#3.3

McDermott, Drew V. (2001). *Mind and Mechanism*. Cambridge, MA: MIT Press.

McKenna, M. (2005). "Where Frankfurt and Strawson Meet." *Midwest Studies in Philosophy: Free Will and Moral Responsibility*, Vol. 29. Ed. Peter A. French and Howard K. Wettstein. Guest Ed. John Martin Fischer. Malden, MA: Blackwell Publishing. 163-180.

McKenna, M. S. (1998). "The Limits of Evil and the Role of Moral Address: A Defense of Strawsonian Compatibilism." *The Journal of Ethics Vol. 2.* Netherlands: Kluwer Academic Publishers. 123-142.

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity.* Cambridge, MA: MIT Press.

Metzinger, T. (2000). "The *Subjectivity* of Subjective Experience: A Representationalist Analysis of the First-Person Perspective." In *Neural Correlates of Consciousness: Empirical and Conceptual Questions*. Ed. Thomas Metzinger. Cambridge, MA: MIT Press. 285-306.

Minsky, M. L. (1965). "Matter, Mind, and Models." *Semantic Information Processing*. Ed. Marvin Minsky. Cambridge, MA: MIT Press. (1968). 425-432.

Nagel, T. (1986). "Freedom." *Oxford Readings in Philosophy: Free Will, 2nd Ed.* Ed. Gary Watson. Oxford: Oxford University Press. pp. 229-256. 2004.

Nozick, R. (1981). *Philosophical Explanations* (Chapter 4 "Free Will," part I "Choice and Indeterminism"). Cambridge, MA: Harvard University Press. 291-316.

O'Connor, T. (1993). "Indeterminism and Free Agency: Three Recent Views." Philosophy and Phenomenological Research Vol. 53, No. 3. 499-526.

Popper, K. R. (1950a). "Indeterminism in Quantum Physics and in Classical Physics. Part I." *The British Journal for the Philosophy of Science*, Vol. 1, No. 2. 117-133.

Popper, K. R. (1950b). "Indeterminism in Quantum Physics and in Classical Physics. Part II." *The British Journal for the Philosophy of Science*, Vol. 1, No. 3. 173-195.

Ramachandran, V. S., Rogers-Ramachandran, D. (1996). "Synaesthesia in Phantom Limbs Induced with Mirrors." *Proceedings: Biological Sciences*, Vol. 263, No. 1369. 377-386.

Russell, P. (2005). "Responsibility and the Condition of Moral Sense." *Philosophical Topics*. Ed. John M. Fischer.

Russell, P. (1992). "Strawson's Way of Naturalizing Responsibility." *Ethics Vol. 102 No. 2*. Ed. John Deigh. Chicago: University of Chicago Press. 287-302.

Scanlon, T. M. (1988). "The Significance of Choice." *Oxford Readings in Philosophy: Free Will, 2nd Ed.* Ed. Gary Watson. Oxford: Oxford University Press. 352-371. 2004.

Searle, J. (2004). *Mind: A Brief Introduction*. New York, NY: Oxford University Press.

Searle, J. (1984). *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.

Silver, D. (2005). "A Strawsonian Defense of Corporate Moral Responsibility." *American Philosophical Quarterly*, Vol. 42, No. 4. 279-293.

Smart, J. J. C. (1961). "Free Will, Praise and Blame." *Oxford Readings in Philosophy: Free Will, 2nd Ed.* Ed. Gary Watson. Oxford: Oxford University Press. 58-71. 2004.

Smilansky, S. (2005). "Free Will and Respect for Persons." *Midwest Studies in Philosophy: Free Will and Moral Responsibility*, Vol. 29. Ed. Peter A. French and Howard K. Wettstein. Guest Ed. John Martin Fischer. Malden, MA: Blackwell Publishing. 248-261.

Smilansky, S. (2000). *Free Will and Illusion*. New York, NY: Clarendon Press.

Strawson, P. F. (1983). *Skepticism and Naturalism: Some Varieties*. New York, NY: Columbia University Press. 1985.

Strawson, P. (1963). "Freedom and Resentment." *Oxford Readings in Philosophy: Free Will, 2nd Ed.* Ed. Gary Watson. Oxford: Oxford University Press. 72-93. 2004.

van Inwagen, P. (1998). "The Mystery of Metaphysical Freedom." *Metaphysics: The Big Questions*. Malden, MA: Blackwell Publishing Ltd. 1998. 365-374.

van Inwagen, P. (1975). "The Incompatibility of Free Will and Determinism." *Agency and Responsibility: Essays on the Metaphysics of Freedom*. Ed. Laura Waddell Ekstorm. Boulder, Colorado: Westview Press. 2001. 17-29.

Wegner, D. M. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Wolf, S. (1981). "The Importance of Free Will." *Perspectives on Moral Responsibility*. Ed. John Martin Fischer and Mark Ravizza. Ithaca, NY: Cornell University Press. 1993. 101-118.

Appendix 1

"The mirror-box. A mirror is placed vertically in the centre of a wooden or cardboard box whose top and front surfaces have been removed. The patient places his normal hand on one side and looks into the mirror. This creates the illusion that the phantom hand has been resurrected" (Ramachandran & Rogers-Ramachandran 1996, 378).[76]

---

[76] Explanation found in: Ramachandran, V. S., Rogers-Ramachandran, D. (1996). "Synaesthesia in Phantom Limbs Induced with Mirrors." *Proceedings: Biological Sciences*, Vol. 263, No. 1369. 377-386.
Image taken from: http://www.tbpiukgroup.homestead.com/files/mirrorbox_3.jpg

Appendix 2

The Ponzo Illusion was first demonstrated by Mario Ponzo, an Italian psychologist, in 1913. The following are two versions of the Ponzo Illusion: in the first image (Figure 1), the bar labelled A appears smaller than the bar labelled B and in the second image (Figure 2), many people will judge the upper figure to be smaller when in fact both curved figures are the same size.[77]

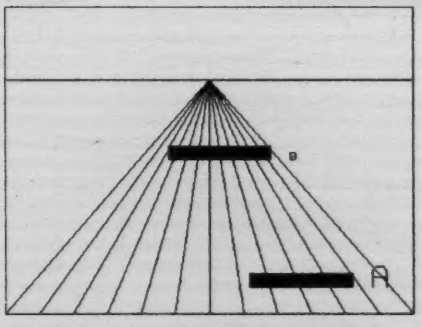

Figure 1

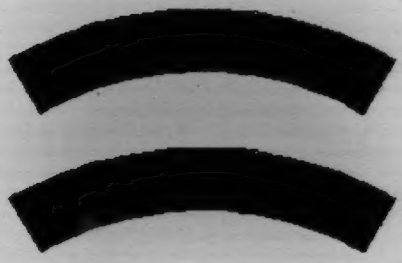

Figure 2

[77] Both images, (Figure 1) and (Figure 2), are taken from Donald E. Simanek's online article "The Moon Illusion, An Unsolved Mystery." http://www.lhup.edu/~dsimanek/3d/moonillu.htm.